University of Southampton

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Assessment of Speech Recognition and Quality of Life as Outcome Measures in Otitis Media with Effusion in Arabic Children

by

Sarah Zohair Alsebai

Thesis for the degree of Doctor of Philosophy

February 2024

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences

Institute of Sound and Vibration Research

Thesis for the degree of Doctor of Philosophy

Assessment of Speech Recognition and Quality of Life as Outcome measures in Otitis Media with Effusion in Arabic children

by

Sarah Zohair Alsebai

Otitis Media with effusion (OME) is the most common cause of conductive hearing loss (CHL) in children and is associated with poor health-related quality of life (HR-QoL). The Paediatric Arabic Auditory Speech Test in Quit (PAAST SiQ) and Arabic OM6 (AOM6) were investigated to assess the use of speech recognition and quality of life (QoL), respectively, as outcome measures in Arabic-speaking children with OME. The main aims were (1) Equalising intelligibility of the PAAST SiQ speech material, (2) Ensuring PAAST SiQ sensitivity to OME-related simulated CHL (SCHL), (3) Developing an Arabic questionnaire to assess QoL in children with OME and (4) Assessing the use of PAAST and AOM6 in Speech recognition and QoL, respectively, in Arabic-speaking children with OME.

The PAAST SiQ was assessed for its homogeneity in Arabic-speaking normal hearing (NH) adults (n=17) using an interleaved adaptive procedure (ILAP). The range of SRTs of the words in the post-equalisation stage (n= 30) was 1.95 dB (±0.98 dB), indicating that the words of the PAAST SiQ were equal in intelligibility. The PAAST SiQ was also assessed for its sensitivity to OME-related simulated conductive hearing loss (SCHL) in Arabic-speaking NH adults (n= 30). The main results were (1) All word SRTs increased significantly with the change in conditions from normal to SCHL (independent t-test against a value of 0 $p < 0.001$) and (2) some words were found to be more sensitive to SCHL.

The AOM6 was developed by translating the OM6 questionnaire to assess QoL in Arabic children with OME. The AOM6 was culturally acceptable, easy to understand, had good internal consistency (Cronbach's alpha (AOM6 Total) = 0.72), no floor and ceiling effect, and partially good construct validity.

The PAAST SiQ and AOM6 were assessed in Arabic-speaking children (aged 3.5-6 years) diagnosed with OME (n = 50) to investigate the effect of OME on speech recognition and QoL, investigate the relationship between speech recognition and QoL, and assess the test-retest reliability of the PAAST SiQ and AOM6. The main results showed that (1) The OME status of the ear led to approximately 5 dB significant increase ($p = 0.006$) in SRTs compared to the No OME status in unilateral OME, (2) The SRTs significantly correlated with Pure-tone audiometry hearing thresholds (PTA-HL) ($p < 0.001$), and it was predicted that a 30 dBA SRT would equate to PTA-HL of approximately 14 dBHL, (3) There was a moderate correlation between SRTs and AOM6 total scores ($r = 0.43$, $p = 0.002$), and (4) The PAAST SiQ and AOM6 had good test re-test reliability.

The current thesis research found the PAAST SiQ to be a reliable test with good discriminative validity, and the AOM6 a reliable measure with suitable psychometric properties, suggesting that both tests are acceptable outcome measure tools for the assessment of speech recognition and QoL, respectively, in Arabic children with OME. Future recommended work includes further validation, application in larger studies, and assessment of the effect of intervention on test scores.

Table of Contents

Table of Tables

Table of Figures

Research Thesis: Declaration of Authorship

Print name: Sarah Zohair Alsebai

Title of thesis: Assessment of Speech Recognition and Quality of Life as Outcome Measures in Otitis Media with Effusion in Arabic children

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 King Abdulaziz University (KAAU) and the Joint Supervision Programme (JSP) for sponsoring me during my PhD, and I would like to especially thank Ms. Omaima Olagy for her support. I would also like to thank my KAAU supervisor Dr. Afaf Bamanie, who was my guide and mentor throughout the years since I started working in the Audiology clinic and during my PhD. I would also like to thank Dr. Mohammad Noman, for his participation in my experiments and for generously carrying the weight of the Audiovestibular Clinic during my PhD. I would also like to thank Dr. Faisal Zawawy and Dr. Ammar Habibullah (KAAU) for their help recruiting participants from the ENT department. A special thank you to Mr. Osama Nassar, for devoting his time to help me recruit participants. Thank you for all the participants, adults and children, who took part in my studies.

I would like to thank all those who supervised me during this journey. Dr. Daniel Rowan, thank you for teaching me to be a better researcher and to question everything and thank you for providing me with test codes. Dr. Hannah Semeraro, thank you for your support and guidance. Dr. Ben Lineton, I will always be thankful for your guidance and support, both academically and technically. Prof. Nicci Campbell, thank you for guiding me and boosting my confidence.

I would like to thank my best friend, Dr. Iman Rawas, I am very lucky that we did our PhD together, I don't how I would've survived this journey without you by my side, thank you for being my person.

I would like to thank my husband Dr. Majed Omar for supporting me during this journey and taking care of our family when I couldn't. I would also like to thank my children. Dana, thank you for being my friend, for cheering for me and for "fixing" my writing mistakes. Deema, your kindness overwhelms me, thank you for being so understanding and helpful. Mohammed and Ahmed, at some point I had to decide to either do my PhD or delay it because you were born right before I got accepted, I'm glad I decided to do it, you both made it a fun journey! I would like to thank my father, thank you for always believing in me and teaching me to never stop learning.

Last but not least, I would like to express the utmost gratitude to my mother, my constant unwavering source of prayers and support, thank you for everything.

I dedicate this thesis to my parents, Mrs. Ilham Telmisani and Prof. Zohair Alsebai, I wouldn't be where I am without you.

Definitions and Abbreviations

Abbreviation	Definition
95%CI	95% confidence intervals
ABR	Auditory brainstem response
AC	Air conduction thresholds in PTA
ALSPAC	Avon Longitudinal Study of Parents and Children
AMTT	Automated McCormick Toy Test
ANOVA	Analysis of Variance
AOM/rAOM	Acute otitis media/ Recurrent acute otitis media
AOM6	Arabic otitis media 6 (OM6) Questionnaire
AOM6 (BaS)/ BaS	Behaviour and symptoms aspect of OM6/AOM6
AOM6 (HaS)/ HaS	Hearing and Speech aspects of the OM6/AOM6
AOM6 (HI)/ HI	Hearing item in the AOM6
AP	Adaptive procedure
AUC	Area under the curve
b	Slope of regression line
BC	Bone conduction thresholds in PTA
CAPT	Chear Auditory Perception Test
CCT	Consonant Confusion Test
CHL	Conductive hearing loss
CIQ	Caregiver impact questionnaire
COM	Chronic otitis media
Condition 2k	OME-related SCHL with a knee point of 2kHz
Condition 4K	OME-related SCHL with a knee point of 4kHz
Condition N	Normal condition (no attenuation)
CSOM	Chronic suppurative otitis media
D/U	Down / Up rule in Adaptive procedure
dB	Decibel
dBA	A-weighted Decibel
DDT	Dichotic digit test
ECLiPS	Evaluation of Children’s Listening and Processing Skills
GIN	Gaps in noise Test
GUI	Graphical user interface
H0	Null hypothesis
HL	Hearing loss
HLS	Hearing loss simulation
HR-QoL	Health related quality of life
IL	Initial level
IL-SRT	The difference between the initial level and the threshold
JCIHS	Joint Committee of Infant Hearing and Screening
L	Location
LMM	Linear Mixed Model
MCS	Monte Carlo simulations

MoCS	Method of constant stimuli
MTT	McCormick Toy Test
n	Sample size
NH	Normal hearing
NU-CHIPS	North-western University Children's Perception of Speech
OM	Otitis media
OM6	Otitis media - 6 questionnaire
OME	Otitis media with effusion
PAAST	Paediatric Arabic Auditory Speech Test
PCA	Principle component analysis
pDev	Goodness of Fit
PedsQL	Pediatric QoL questionnaire
PF	Psychometric Function
PSI	Paediatric Speech Intelligibility
PTA	Pure-Tone Audiometry
PTA-HL	Pure-Tone Audiometry hearing threshold levels
QoL	Quality of Life
r	Correlation coefficient
R^2	Goodness-of-fit in linear regression
RCT	Randomised clinical trial
REML	Restricted maximum likelihood
RM-ANOVA	Repeated measure – Analysis of Variance
ROC	Receiver operating characteristic
RTS	Reception threshold for sentences
S	Slope
SCHL	Simulated conductive hearing loss
SD	Standard deviation
SDω	Within-subject standard deviation
SE	Standard error
SiN	Speech recognition in noise
SiQ	Speech recognition in quiet
SLM	Sound level meter
SNHL	Sensorineural hearing Loss
SNR	Speech-in-noise ratio
SRA	Sentence recall accuracy
SRE	Studentized residuals
SRM	Standardized response mean
SRT	Speech recognition threshold
VAS	Visual analogue scale
V^{Diff}	Difference between the estimated threshold and the true threshold
$V^{Estimate}$	Value of the estimated SRT
V^{True}	Value of the true SRT
WDT	Word discrimination threshold
WIPI	Word Intelligibility by Picture Identification

Chapter 1 Introduction

Otitis media with effusion (OME) is defined as the presence of a middle ear effusion in the absence of infection (Rosenfeld *et al.*, 2022), and it is one of the most common conditions in childhood. In their review, Rosenfeld *et al.* (2022) commented that OME could be considered an 'occupational hazard of early childhood'. Therefore, good assessment, management, and follow-up with the appropriate outcome measure tools are required to ensure good quality care for children with OME.

One of the most common morbidities of OME is conductive hearing loss (CHL), which consequently can lead to the child's inability to recognise speech, negatively affecting their vocabulary acquisition and speech development. Factors including the chronicity of disease, hearing loss, and possibly affected communication abilities can lead to a decline in academic learning and behavioural issues. All these issues can collectively decrease the quality of life (QoL) in children with OME. According to a recent review, it was recommended to include hearing assessment and measurement of quality of life (QoL) as outcome measures in children with OME (Homøe *et al.*, 2020).

This PhD aimed to assess speech recognition and QoL as outcome measures in Arabic-speaking children with OME. The outcome measures currently used in Arabic-speaking children with OME are clinical examination of the ear, tympanometry, and conventional hearing assessments such as Pure Tone Audiometry (PTA) (Al-humaid *et al*., 2014). One limitation of PTA is that it may not be suitable to perform for many 2.5 – 4 year-olds, despite that age group being more commonly affected by OME than older children (Martines *et al*., 2011). Additionally, PTA does not test speech recognition, and it has been argued that doing so is important especially in research (Hall *et al.,* 2007). The aforementioned outcome measures do not directly provide information on quality of life (QoL). Consequently, it is unclear how OME affects speech recognition and QoL in Arabic-speaking children and how the treatments for OME would affect these measures in that population.

The particular interest in Arabic-speaking children stems from the fact that tools that assess speech recognition and QoL are language specific. There are currently a limited number of automated validated Arabic speech tests, and none are suitable for children as young as 2.5-4 years. Also, to date, there are no Arabic disease-specific QoL questionnaires for children with OME.

Speech recognition tests can assess hearing in children aged 2.5 years and older, especially if the utilised test is easy and can be automated and readily available on smart devices, such as the Automated McCormick toy test (AMTT) (Hall *et al.*, 2007). An Arabic version of the AMTT, the Paediatric Arabic Auditory Speech Test (PAAST), was developed by a previous PhD student at the University of Southampton, Rania AL-Kahtani (Al-Kahtani, 2020), to assess the speech recognition in noise (SiN) in Arabic-speaking children with sensorineural hearing loss (SNHL). The PAAST was developed to address the lack of automated Arabic speech recognition tests for children. In the current PhD, the PAAST in quiet (SiQ) was researched for its validity and ability to detect conductive hearing loss (CHL) in children with OME.

The health-related quality of life (HR-QoL) is another important aspect when assessing children with OME. The OM6 is a QoL questionnaire that was developed by Rosenfeld *et al*. (1997), it is a valid, reliable, and easy tool, and it is one of the most commonly used QoL questionnaires in randomised clinical trials (RCTs) (Gan *et al.*, 2018). One of the aims of this PhD was to develop an Arabic QoL tool by translating and culturally adapting the OM6 from English to Arabic and testing the reliability and validity of the Arabic OM6 (AOM6).

The motivation behind this PhD was to improve the outcome measure tools in children with otitis media with effusion (OME) in Saudi Arabia by assessing PAAST SiQ and AOM6 as tools to investigate the use of speech recognition and QoL as outcome measures in Arabic-speaking children with OME.

The broad aims of this PhD are:

Equalise the intelligibility of the speech material of the Paediatric Arabic Auditory Speech Test in quiet (PAAST SiQ) and determine its sensitivity to conductive hearing loss.

Develop an Arabic disease-specific QoL tool for Arabic children with OME by translating the English OM6 questionnaire and validating the Arabic version.

Explore the effects of OME on speech recognition and QoL, using the PAAST and AOM6, respectively.

1.1 Structure of the thesis

This research project consisted of three main parts (illustrated in Figure 1.1). The first part consisted of two studies (Experiment 1 (A and B) and 2), which focused on equalising the intelligibility of the PAAST SiQ and measuring its sensitivity to OME-related simulated conductive hearing loss (SCHL). The findings of these studies indicated that the PAAST SiQ was appropriate to explore its use in assessing speech recognition in Arabic children with OME. The second part consisted of one study (Experiment 3), which was directed toward developing the Arabic version of the OM6 (AOM6). The third part was the final study that aimed to answer the main research question of this PhD, which was: "Can speech recognition and Quality of life measured by the PAAST SiQ and AOM6, respectively, be considered good outcome measures in Arabic-speaking children with OME?".

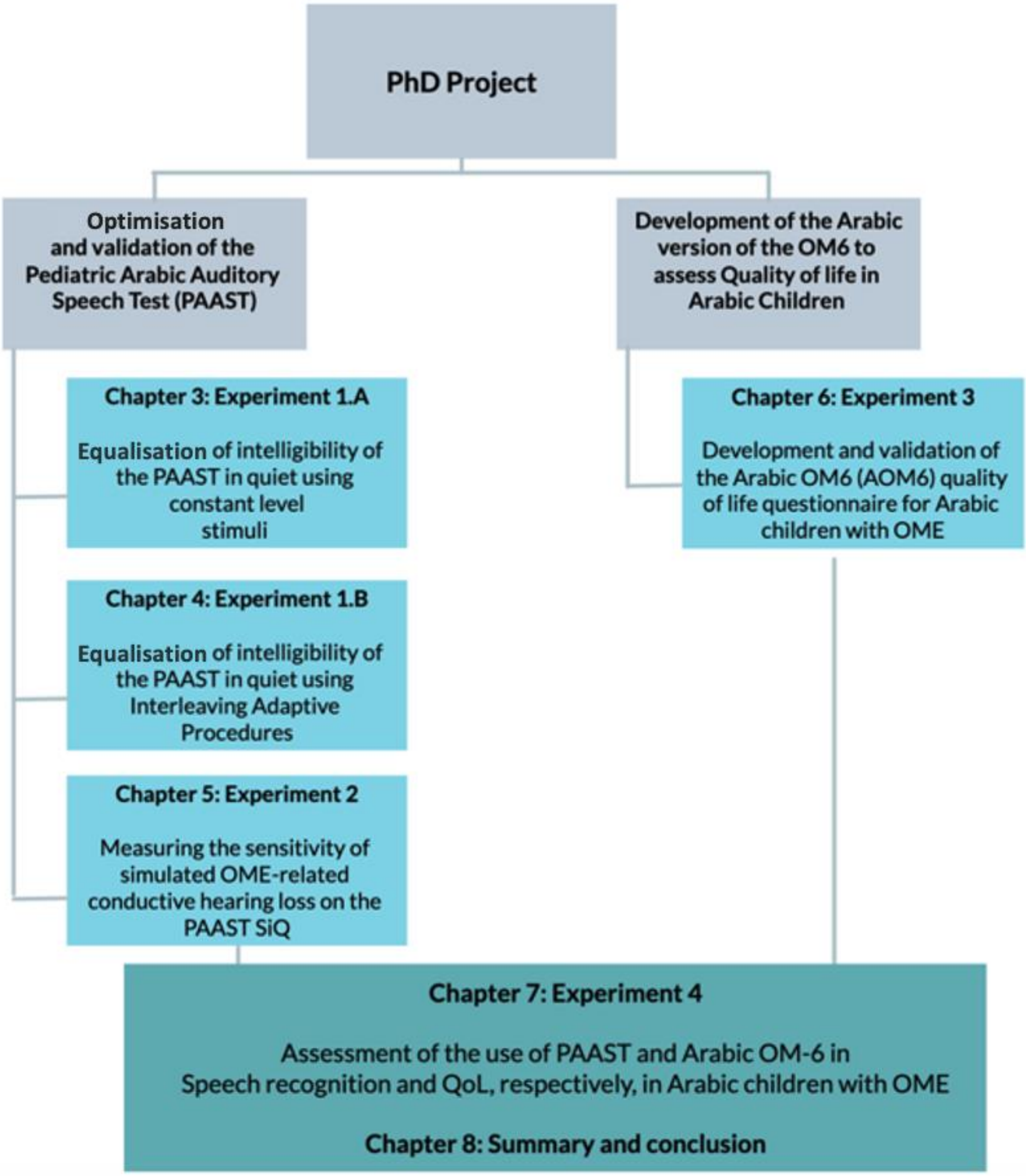


Figure 1.1 A flow chart of the studies conducted in the current PhD thesis

The current thesis was structured in chapters as follows:

Chapter 2 discussed the literature review of OME, the situation of OME in children in Saudi Arabia, and an overview of the outcome measures in OME. It also discussed the effect of OME on the speech recognition and quality of life in literature. Towards the end of the chapter, the gaps in knowledge and the research aims and objectives of this PhD were discussed.

Chapter 3 was composed of the pilot study that originally aimed to assess the sensitivity of the PAAST SIQ to SCHL in normal hearing (NH) adults, and Experiment 1.A which aimed to assess the homogeneity of the words of the PAAST SiQ using the method of constant stimuli (MoCS), also in NH adults.

Chapter 4 covered Experiment 1.B which aimed to equalise the intelligibility of the PAAST SIQ using the interleaved adaptive procedure (ILAP) in NH adults, that led to the final, equalised for intelligibility version of the PAAST SIQ. This chapter also discussed the Monte Carlo simulation (MCS) study, which assessed the parameters of the PAAST SiQ for their accuracy and precision.

Chapter 5 covered the post-equalisation assessment of the homogeneity of the words of the PAAST SiQ, and Experiment 2 which aimed to measure the effect of OME-related SCHL on the PAAST SIQ and its words in NH adults.

Chapter 6 discussed Experiment 3, which aimed to develop an Arabic QoL questionnaire for Arabic children with OME. The OM6 questionnaire was translated and cross-culturally adapted to Arabic, resulting in the AOM6. Psychometric properties, including aspects of reliability and construct validity, as well as floor and ceiling effects, were assessed by applying the questionnaire to parents of Arabic-speaking children with OME.

Chapter 7 covered Experiment 4, which aimed to answer the main question of this PhD, as to whether the PAAST SiQ and AOM6 can be considered appropriate measure tools for assessing the effect of OME on speech recognition and QoL in Arabic-speaking children with OME. Objectives of this study included assessing the effect of OME on the aforementioned tools, exploring the relationship between speech recognition and QoL, and measuring the test-retest reliability of the PAAST SiQ and AOM6.

Chapter 8 summarised the work done in the current PhD and provided conclusion and recommendations for future work.

1.2 Contributions to knowledge

The main contributions from the current thesis to knowledge and research community are as follows:

- The validation of the PAAST SiQ, an automated and easy-to-use tool (both by parents and children (Al-Kahtani, 2020)) which can be implemented as a smart device application. The PAAST SiQ has shown a good positive correlation with PTA average thresholds ($p < 0.001$) in children. These characteristics allowed the PAAST SiQ to be a potentially helpful tool in assessing hearing in Arabic children aged 3.5 years and older with OME in remote areas where conventional or portable PTA devices are unavailable. The PAAST SiQ can also be a valuable tool in research to study the development of speech recognition in Arabic-speaking children and the effect of treatment on speech recognition.
- The AOM6 questionnaire is, to date, the first disease-specific Arabic QoL questionnaire for children with OME, which has been tested for a number of psychometric properties. The AOM6 can potentially be used in assessing QoL in Arabic children with OME, but preferably studies with large sample sizes and additional elements, including pre- and post-treatment data to assess responsiveness, would ensure the tool's validity.
- The current PhD explored the relationship between speech recognition and QoL, which has not been assessed in previous studies, despite the closeness of these measures to each other.

Research activities during the current PhD include:

- Oral Presentation on ‘Measuring speech recognition and quality of life in Arabic children with OME using the PAAST SiQ and Arabic OM-6 questionnaire’, Hearing, Audio and Audiological Sciences (HAAS) Meeting, Southampton, July 2022.
- Oral presentation and poster on ‘Paediatric Arabic Auditory Speech Test (PAAST) in Quiet as a tool to detecting conductive hearing loss’, HSG Research Away Day, Southampton, UK, July 2019.

Chapter 2 Background and literature review

This chapter reviewed otitis media with effusion (OME) in general, its complications, its outcome measures, and the outcomes assessed in children in Saudi Arabia. The outcome measures of interest in the current thesis were speech recognition and QoL, which were reviewed in this chapter, along with the proposed tools to assess them.

The gaps in knowledge, research question, and aims of the current PhD project were discussed at the end of the chapter.

2.1 Otitis media with effusion

2.1.1 Epidemiology, aetiology, diagnosis, and management

Otitis media with effusion (OME) is defined as the presence of a middle ear effusion in the absence of infection (Atkinson *et al.,* 2015). Approximately 80% of all children younger than four years old have had at least one episode of OME (Saber *et al.*, 2021). Acute otitis media (AOM) is one of the most common conditions in childhood and is one of the leading causes of OME. Most children have experienced at least one AOM episode by age three years, and by age six years, nearly 40% have experienced three or more episodes (Rosenfeld *et al.*, 2022). Causes of how OME can develop are illustrated in Figure 2.1.

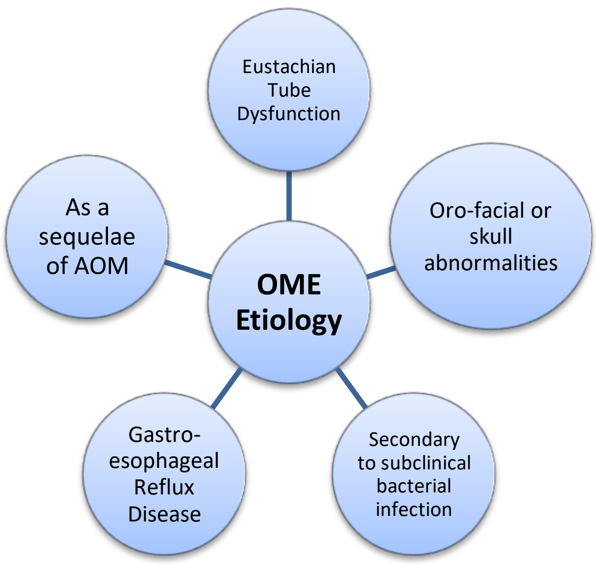


Figure 2.1 Etiology of OME development (adapted from Atkinson *et al.,* 2015)

Children younger than seven years are at increased risk of otitis media (OM) because of physiological and anatomical factors, including their immature immune systems and poor function of the eustachian tube (Rosenfeld *et al.*, 2022). The Eustachian tube, a slender connection between the middle ear and nasopharynx that equalises the pressure in the middle ear cleft, is shorter and more horizontal in children than in adults. This anatomical difference can lead to eustachian tube dysfunction, which in turn may obstruct inflammatory exudates in the middle ear secondary to infection or allergy, leading to AOM or OME, or both (Rosenfeld *et al.*, 2022).

Middle ear effusion can be present in both AOM and OME (sometimes called chronic otitis media (COM)). Acute otitis media is a condition characterised by the rapid onset of acute symptoms of middle ear inflammation, often combined with the presence of OME. On the other hand, OME/COM is a condition causing mild CHL due to the presence of effusion in the middle ear behind an intact tympanic membrane in the absence of active inflammation of the Tympanic membrane (TM) (A. A. Timmerman, 2008; Rosenfeld *et al.*, 2022).

Another form of middle ear infection is chronic suppurative otitis media (CSOM), which is a chronic inflammation without intact TM, and is often associated with discharge from the ear (otorrhea) (Rosarioa and Mendez, 2023). Chronic suppurative otitis media is not the main topic of discussion in the current PhD, but some studies mentioned in this review discussed this type of infection.

The most common morbidity of OME is conductive hearing loss (CHL), especially when OME is persistent or recurrent, which can lead to delayed speech and language development in children as well as adverse effects on behaviour linked to poor listening skills (Homøe *et al.*, 2020).

The commonality of the condition and the ensuing complications on the general health status and speech and language abilities (discussed in detail in Section 2.2.7) warrants thorough assessment, management, and follow-up with the appropriate outcome measure tools to ensure good quality care for the child. Additionally, it is vital to ensure reliable outcome measures due to the nature of the disease. Otitis media with effusion is a chronic disease with self-limiting nature, necessitating a comprehensive set of reliable outcome measures to facilitate decisions regarding intervention (Chessman *et al.*, 2016). Figure 2.2 illustrates the process of diagnosis and management of OME proposed by the American Academy of Otolaryngology (Rosenfeld et al., 2016). In Saudi Arabia, the management of OME follows the American guidelines.

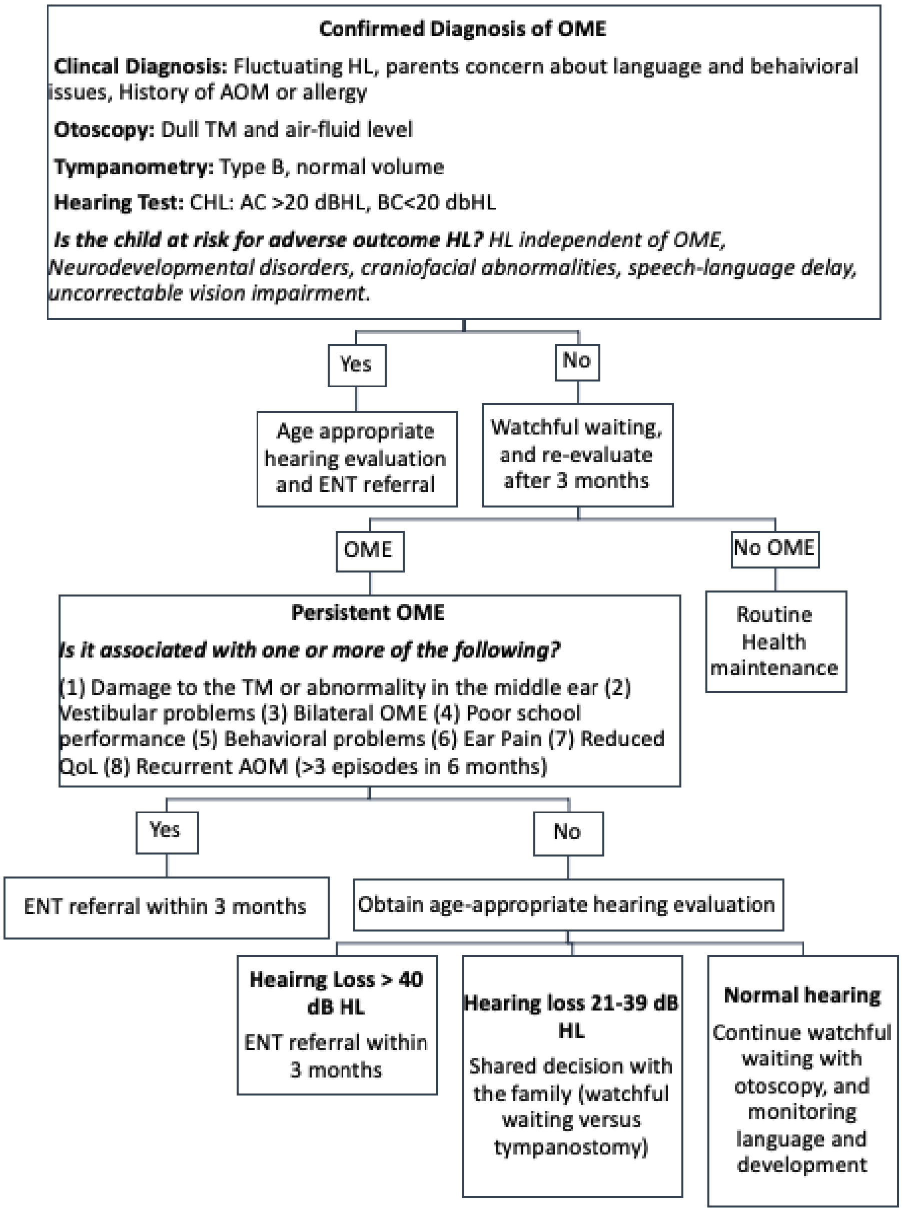


Figure 2.2 Diagnosis and management of OME in children (adapted from Rosenfeld et al., 2016)

According to Figure 2.2, hearing loss (HL) assessment is necessary for deciding the management path for a child with OME. The American Academy of Otolaryngology recommends assessing hearing by an age-appropriate test, and most commonly, Pure-tone Audiometry (PTA) is the test of choice. It is also recommended that if a speech test is considered, the test should be in quiet (Rosenfeld *et al*., 2016), given that it can then be used to predict hearing thresholds (Hall et al., 2007). The importance of assessment of hearing, especially in persistent or recurrent OME is highlighted in the NICE guidelines for managing OME and making decisions regarding intervention (NICE, 2018). Therefore, choosing an appropriate-for-age, readily available hearing test is crucial for managing OME. Quality of life is an important aspect to consider clinically to make decisions about treatment and measure the response to the management chosen, and in research to assess the effectiveness of a treatment.

2.1.2 Outcome measures of treatment of otitis media with effusion

There are two aspects to the term outcome measures, health outcome and outcome measures. Health outcome can be defined as “a change in the health status of an individual, group or population which is attributable to a planned intervention or series of interventions” (World Health Organization 1998, p10.). Outcome measures evaluate the effectiveness of a particular intervention (Chessman *et al*., 2016), and choosing suitable outcome measures allows for the delivery of safe and effective management plans and interventions (AHP, 2019), thus improving learning and care over time (Pantaleon, 2019). Other uses for outcome measures are (AHP, 2019):

- Identifying meaningful change for the person accessing services (e.g., quality of life).
- Demonstrating the impact and value of services (e.g., to patients and practitioners).
- Identifying areas for improvement.
- Benchmarking against other organisations/services/standards.

Three main aspects allow for assessment of the appropriateness of an outcome measure (AHP, 2019):

1. Initial considerations: This includes identifying the outcomes that would represent the change in the health status and then identifying the measures that best evaluate them. An example would be identifying that hearing is one of the health outcomes of intervention in OME. Therefore, a hearing test is best suited to measure this outcome.

2. Acceptability and utility: Focusing on whether the outcome measure is user-friendly and relevant, and assessing its feasibility within the practice. An example of this would be ensuring that the QoL questionnaire used is culturally acceptable and easy to understand.

3. Measurement properties: Outlining key measurement properties, including validity, reliability, and responsiveness/sensitivity to change. An Example of this is assessing the psychometric properties of a QoL questionnaire.

Measuring treatment outcomes in a disease such as OME is essential, given the disease’s high possibility of natural resolution, to determine whether the resolution of OME is spontaneous or due to an intervention (Rosenfeld, 2003). In a literature review on the outcome measures of treatment of OME in 171 studies, various outcome measures were reported (Figure 2.3) (Chessman *et al*., 2016). The review found that the most investigated outcome measure was the resolution of OME based on clinical examination using otoscopy, followed by the hearing level, for which most studies used PTA. Speech perception as an outcome measure was investigated in four studies. Another frequently used outcome measure was HR-QoL, where some studies used a well-defined validated disease-specific questionnaire, such as the OM6 questionnaire, while others used generic or study-specific questionnaires that were not validated.

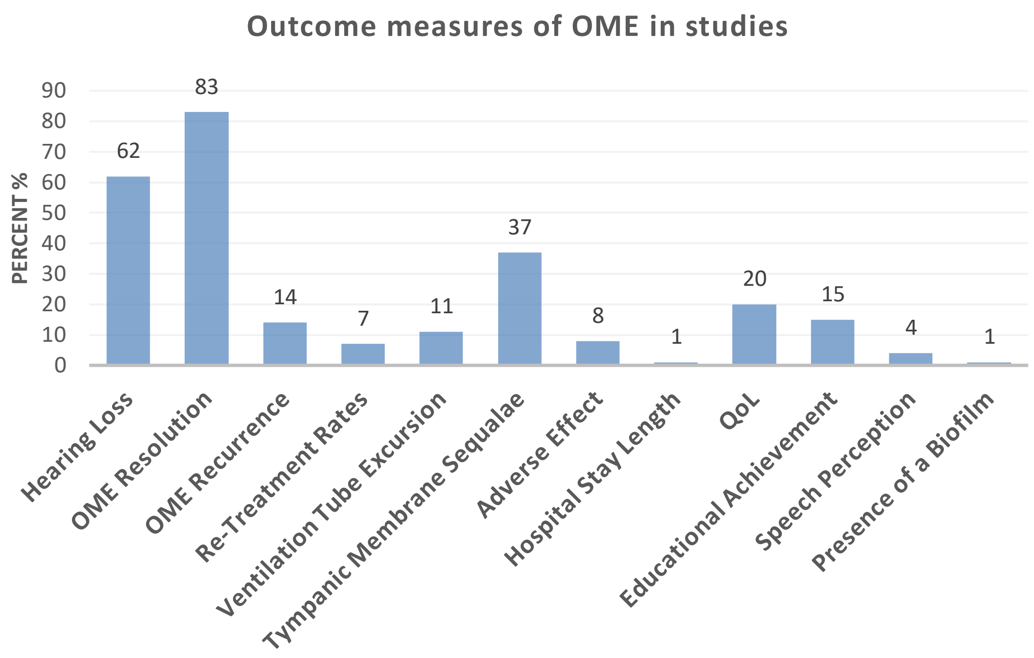


Figure 2.3 Percentage of outcome measures used in studies on OME in children. (Adapted from Chessman *et al.,* 2016)

Speech recognition testing is a measure that is useful both clinically and in research. Speech tests in quiet such as the McCormick toy test (MTT), which is a valid, reliable, and easy test, can be used to predict hearing thresholds in children 2.5 years and older (Hall *et al.,* 2007) (discussed in details in Section 2.2.4.2). Auditory abilities such as perceptions of sounds (PTA), as well as developmental aspects of speech discrimination and recognition, are of interest in research to understand the effect of OME on children's speech development (Hall *et al.,* 2007). Considering there are no automated speech recognition tests for children as young as 2.5 years in Arabic (Garadat et al., 2017), this PhD aimed to validate an automated speech recognition test for children. The validation of the Arabic version of the MTT was important because translating a test is not enough to ensure that it fulfils all aspects of the feasibility and validity of an outcome measure tool.

Quality of life is another important outcome measure in children with OME. The overall persistence and severity of this disease can impact QoL. Therefore, assessment of QoL can add value to the diagnostic process and the evaluation of treatment effects (Homøe *et al.*, 2020). There are several disease-specific QoL questionnaires in OME, which are characterised by inquiring about specific elements, including the physical and psychological wellbeing of the child, as well as their hearing and speech status (Homøe *et al.*, 2020). There is a lack of Arabic-language tools that measure disease-specific QoL in children with OME in the Arab world. The need for a validated Arabic QoL assessment tool motivated this PhD's aim to develop such a tool.

In order to use an outcome measure in a clinical setting or a randomized clinical trial (RCT) designed to assess the effectiveness of a particular treatment, it is crucial to ensure the measure's validity (Liu *et al.*, 2020). Much of the current project was dedicated to ensuring the validity of the tools intended to assess speech recognition in quiet (SiQ) and QoL in Arabic.

2.1.3 Otitis media with effusion in Saudi Arabia

The focus of the current PhD project is children with OME in Saudi Arabia since OME is considered one of the most common childhood morbidities in this area (WHO, 2004), and specific outcome measure tools are unavailable, including speech recognition testing and QoL, that are useful both clinically and in research in Saudi Arabia.

One of the largest studies on the prevalence of HL in Saudi Arabia among children between the ages of 4 and 15 years showed that 13% had HL and that 10% of those children had CHL, while 3% had SNHL and mixed hearing loss (Al-Abduljawad and Zakzouk, 2003). In a study by Al-Muhaimeed (1996), a random survey of 6,421 children aged 0 to 12 years was conducted, 1256 (19.6%) of the total sample were found to be 'at risk' for hearing impairment. Children who had one or more risk factors identified by the Joint Committee of Infant Hearing and Screening (JCIHS) (1982) (mentioned in Al-Muhaimeed, (1996)) were considered 'at risk' of hearing impairment. Children with hearing impairment in this study were found to be 494, representing 39.3% of the 'at risk' children. Of the 494 Hearing impaired children, 326 (66%) had CHL, mostly associated with OME (n = 232). A study on the prevalence and aetiology of hearing loss in Saudi Arabia reported that among n = 2574 children aged 4-8 years, 45 children were diagnosed with hearing loss, among which 84.4% had CHL, with the leading cause being middle ear disease, mostly OME (34.9%) (Al-Rowaily *et al.*, 2012). A recent study by Alkahtani *et al.*, (2019) shed light on the age of identification of hearing loss in children in Saudi Arabia, showing that a total of 533 children who visited the Audiology clinic (total n = 1166 aged 0 -10 years) had hearing loss, 193 of which had CHL with a mean age of identification of 3.9 (2.7) years. This study suggested that there was an issue with detecting hearing loss in general in children in Saudi Arabia, possibly related to several factors, including the lack of hearing assessment tools that can be easily applied to very young children and the lack of awareness of parents on possible signs of OME and associated CHL (Alkahtani *et al.*, 2019).

Several studies were conducted on the prevalence, diagnosis, and management outcomes of OM and its subtypes: AOM, OME, and CSOM in Saudi Arabia (Table 2.1). These studies showed that the diagnosis of OME in Saudi Arabia mainly relied on otoscopy and tympanometry. In one study, diagnosis depended on otoscopy alone (Al-Quaiz, 2001). Hearing tests were performed in most studies, including PTA as a behavioural test and Auditory Brainstem Response (ABR) as an objective test for younger children. Alqahtani *et al.* (2017) attempted to study the effect of OME on the school performance of children assessed by their teachers using a 3-item questionnaire (active – lethargy – lazy) with a yes or no response, but no significant difference was found between children with OME and no OME ($p = 0.067$), suggesting that a detailed assessment using a disease-specific QoL questionnaire could provide better information on how the child is affected by OME. No speech recognition tests or QoL tools were used in these studies. The importance of these measures is discussed in the following sections.

Table 2.1 Outcome Measures used to assess Otitis Media with Effusion in Saudi Arabia

Study	**Participants**	**Tests**	**Results**
Epidemiology of chronic suppurative otitis media in Saudi children (Muhaimeid, Zakzouk and Bafaqeeh, 1993)	N= 6421 (survey) Age: 2 months to 12 years	• Otoscopy • PTA	- 94 children (1.5%) had chronic ear disease and CHL - OME accounted for 66% HL in KSA
Correlates of Various Presentation Modes of Acute Otitis Media in Saudi Children (Al-Quaiz, 2001)	n=140 (AOM) Age ≥ 5 years	• Otoscopy	Majority of AOM 12-30 months
Epidemiology of acute otitis media among Saudi children (Zakzouk, Jamal and Daghistani, 2002)	n= 9540 children age≥12 years	• Tympanometry • otoscopy • PTA	- AOM: 1.05% - Incidence higher up to 4-years-old and lower in the age group 8–12 years. - AOM more in related parents (P>0.001).
Point prevalence of type B tympanogram in Riyadh (Zakzouk and Abduljawad, 2002)	n = 4214 (survey) age 1-8 years	Tympanometry	-Unilateral Type B= 5.7% and -Bilateral Type B= 8.1%
Hearing impairments among Saudi preschool children (Al-Rowaily *et al.*, 2012)	n = 2574 (kindergarten entry) n = 2204 (primary school entry) Age 4-8 Mostly females	• Tympanometry • PTA	45 children were diagnosed with hearing impairment CHL: 84.4% SNHL 15.6% OME: 1.75% First as a cause of HL: OME (34.9%)
Management of Otitis Media with Effusion (Ashoor and Fuer, 2013)	48 children with OME mean age: 6 years (1-16 years)	• Otoscopy • Tympanogram • PTA • ABR	-Acute OME=38: -Chronic OME=10 -Positive correlation in tympanometry and hearing after improvement
Prevalence and risk factors of Otitis Media with effusion in school children in Qassim Region of Saudi Arabia (Humaid *et al.*, 2014)	n=1488 6-12 years	• Otoscopy • Tympanometry • PTA	Prevalence of OME in the study population was 7.5%
Otitis Media in Children at Riyadh Capital City of KSA (Alqahtani *et al.*, 2017)	cross-sectional study n = 1500 Age range 7 - 14	• Otoscopy • Tympanometry • Questionnaire (risk factors and level of performance)	-OME = 10% (150/1500) -No significant difference in School performance between OME and normal children (p = 0.067)

2.2 Speech recognition in children

2.2.1 Audiological testing in children

Generally, hearing in children can be evaluated using different techniques, including visual reinforcement audiometry (VRA), play audiometry, standard PTA, or objective tests like the ABR (Sabo *et al.*, 2003). Choosing the test type and environment depends on the child's age, and the aspect of hearing aimed to be tested, to ensure valid results (Sabo *et al.*, 2003). Appendix A includes the main audiological tests for screening, diagnosis and follow-up in children.

The aim of performing speech tests is to gain information on the perception of speech in children, especially those with OME, given that the periods of hearing deprivation associated with the condition could lead to delays in language development (Jamal *et al.,* 2022). Speech recognition tests can also be used as an easy reliable test to estimate hearing thresholds in children as young as 2.5 years old, who are too old to perform VRA and perhaps too young to understand the concept of play audiometry (McCormick, 1977; Harries and Williamson, 2000; Haggard, 2004; Lovett *et* al.,2013). To choose the appropriate speech test for the child's age, it is essential to understand the levels of audiological skills development in children.

2.2.2 Speech testing in children

Although PTA provides valuable information about the degree and type of hearing loss, auditory speech tests provide insight into auditory skills that could be affected by hearing loss. Speech tests are an essential component of the audiological test battery, as they provide information regarding the sensitivity to the speech stimuli and understanding speech at suprathreshold levels (Mendel, 2008). They are also useful for:

- Examination of speech processing abilities throughout the auditory system (e.g. speech in noise tests in SNHL) and allowing for better understanding of how well the child can perceive speech (Neumann *et al*., 2012).
- Cross-checking the validity of PTA thresholds.
- Use as a prognostic tool of the children's speech, language, reading and cognitive abilities.
- Making decisions about amplification with hearing aids, cochlear implants, language learning methods, and additional audiological rehabilitation.
- Monitoring the child's progress.
- Application in quiet as an alternative to PTA, especially if it was an easy, reliable test that can predict the PTA thresholds (Summerfield *et al.*, 1994).

Speech tests assess the ability of the child to detect, recognise, and understand simple and complex speech stimuli (Singleton and Waltzman, 2015, Vickers *et al.*, 2017), and they should be chosen based on the auditory skill intended to test (Ondáš *et al.*, 2020). A hierarchy of auditory skill development was suggested and tested by Erber ((Erber (1982), cited in Perigoe and Paterson, 2013)) (Figure 2.4).

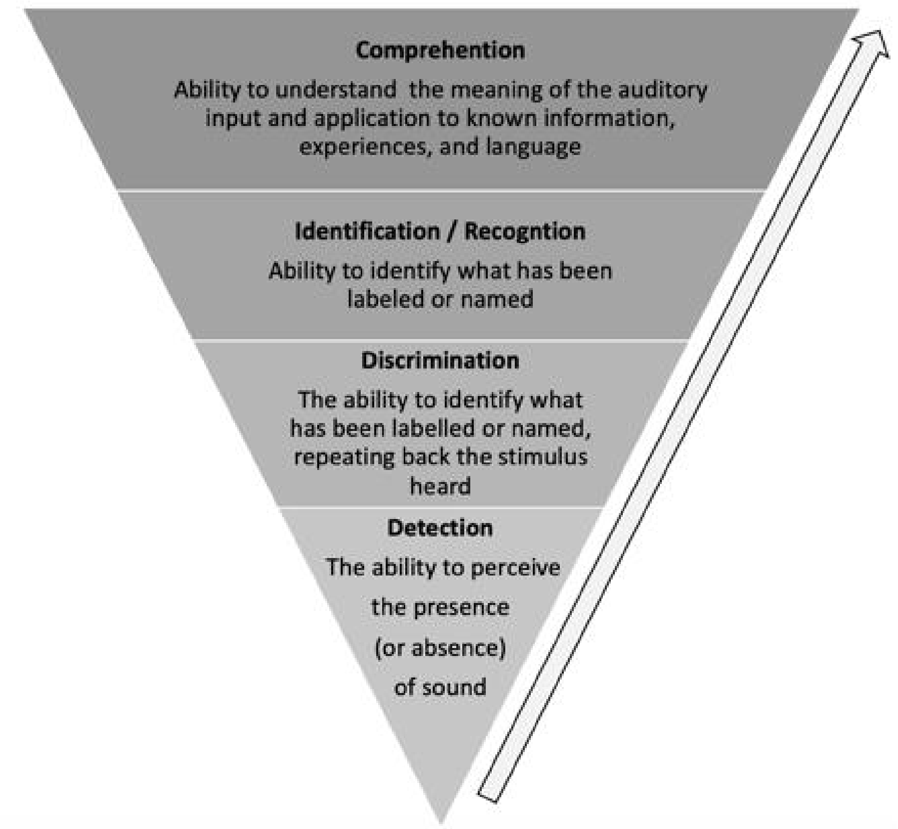


Figure 2.4 Erber's Hierarchy for speech development in children (Adapted from Erber (1982), cited in Perigoe and Paterson, (2013))

Although the levels of development are presented as a hierarchy, they do, in fact, overlap. A normal hearing (NH) child or a child with hearing loss (HL) could be developing all four levels of skills—detection, discrimination, identification /recognition, and comprehension—at the phoneme level, word level, and sentence level concurrently (Perigoe and Paterson, 2013).

Since the current PhD focuses on speech testing in children who are 2.5 years old and older, it was essential to know the audiological development of children at this age. At 24-36 months, a NH child is expected to develop auditory identification or recognition (Perigoe and Paterson, 2013), meaning they should identify a sound and share that identification with another person.

Speech recognition can be tested through several tests, but the most commonly validated test in children is the McCormick toy test (MTT), a behavioural hearing test designed and described by McCormick in 1977 (McCormick, 1977).

The MTT was designed for hearing screening, estimating the hearing threshold, assessing the functional impact of OME and SNHL, and assessing the benefits of hearing aids in children (Hall et al., 2007; Cullington *et al.*, 2017; Brown et al., 2022). The MTT is based on presenting the child with a word via a speaker or headphones and then asking them to choose the picture corresponding to the word they heard. The measure obtained from the MTT is the word discrimination threshold (WDT) or speech recognition threshold (SRT). Details of the MTT are explained in Section 2.2.4.

Other tests that assess speech recognition in children include the Consonant Confusion Test (CCT) and Chear Auditory Perception Test (CAPT) (Vickers *et al.*, 2017). Like the MTT, the CCT is used to assess children aged 2 – 4 years, and the CAPT can be applied as a follow-up to the CCT (Vickers *et al.*, 2017). The CCT is made up of 10 groups of words. Each group is composed of four phonemically similar (same vowel) words. One of the four words in each group is presented to the child in a recorded audio form, and the child is asked to choose the corresponding picture from the four choice pictures displayed in front of them. The CAPT is similar to CCT, except the former comprises advanced words. The measure obtained using the CCT and the CAPT is percent correct, which indicates the percent of correctly recognised words from pictures at a certain level (e.g., 30 dB). Like the MTT, the CCT and CAPT were assessed for reliability and repeatability.

The Word Intelligibility by Picture Identification (WIPI) is one of the widely used American tests for children aged 2.5 years and older (Cienkowski *et al.*, 2009; Flaherty *et al.*, 2022), and similar to CCT and CAPT, the measure obtained from the test is the percent correct at a certain presentation level. Other closed-set tests picture-pointing tests suitable for children aged three and above include North-western University Children's Perception of Speech (NU-CHIPS) and Paediatric Speech Intelligibility (PSI) (Singleton and Waltzman, 2015).

The Arabic version of the MTT was chosen as a potential outcome measure tool in children with OME because the automated MTT (AMTT) was extensively researched for its validity and reliability (Summerfield *et al.*, 1994; Lovett *et al.*, 2013). It was also researched for its use in assessing speech recognition in children with OME (Haggard, 2004; Hall *et al.*, 2007), allowing for comparison in a disease-specific context.

Although the MTT is considered a speech discrimination test (McCormick, 1977), the parameter used for this test is sometimes referred to as the speech recognition threshold (Hall *et al.*, 2007; Al-Kahtani, 2020). The child must discriminate the words to recognise them (Houston *et al.*, 2016) in order to perform the test and obtain the speech recognition thresholds (Hall *et al*., 2007). Therefore, studies differed in referring to the terminology used for the measure of the MTT, where some studies refer to the 71% threshold as the word discrimination threshold (WDT) (Brown *et al*., 2022) and others referred to the same threshold as word recognition threshold (WRT) (Hall *et al*., 2007). Throughout the current PhD, the term speech recognition threshold (SRT) was used to describe the threshold at which a percentage of words were "recognised" using the Arabic version of the AMTT, namely the Paediatric Arabic Auditory Speech Test (PAAST).

Speech recognition testing in children and adults can be performed under different conditions, such as in noise and in quiet, and each testing condition can provide different information about hearing abilities. The following sections discuss the effect of OME on pure-tone thresholds and speech recognition thresholds in different conditions.

2.2.3 OME and its effect on hearing and speech recognition

2.2.3.1 Effect of OME on PTA thresholds and common audiometric configurations

Hearing loss can be categorised according to two main aspects: type and degree. There are three main types of HL. The first is SNHL, caused by a pathology in the inner ear and/or the cochlear nerve, characterised by elevated pure-tone thresholds and an air-bone gap of less than 5 dB. The second type is CHL, caused by a pathology in the outer and/or middle ear, such as OME, and is characterised by normal bone conduction (BC) thresholds (BC ≤15dB HL). The third type is mixed HL; caused by a combined involvement of the outer/middle ear and the inner ear/cochlear nerve, and characterised by a BC threshold of more than 20 dB HL and an air-bone gap equal or greater than 15 dB (Newton, 2009). The degrees of HL are listed in Table 2.2 (Clark, 1981).

Table 2.2 Scale of hearing impairment Adapted from Clark (1981)

Average hearing threshold (dBHL)	**Hearing loss label**
-10 - 15	Normal hearing
16 - 25	Slight HL
26 - 40	Mild HL
41 - 55	Moderate HL
56 - 70	Moderately severe HL
71 – 90	Severe HL
≥ 91	Profound HL

Otitis media with effusion in children can cause CHL of different degrees, ranging from mild to moderate HL, with an average PTA threshold (across frequencies 0.5, 1, and 2 kHz) of 18-35 dB HL (Cai and McPherson, 2017). The most common audiometric configuration in children with CHL caused by OME is rising low frequency HL (Newton, 2009; Silman and Silverman, 1997; Cai and McPherson, 2017). A rising low frequency HL caused by OME is characterised by air-conduction (AC) thresholds >15 dBHL in the low frequencies below a particular cut-off frequency, commonly the 2 kHz (Cai and McPherson, 2017), with a normal BC and large air-bone gap in the low frequencies. The AC threshold decreases as it approaches the cut-off frequency until it reaches normal levels and then plateaus (or decreases) beyond that point (Figure 2.5).

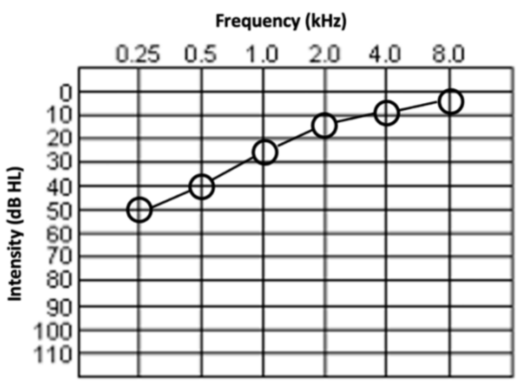


Figure 2.5 Audiometric configuration of CHL caused by OME showing a rising low frequency hearing loss (adapted from Silman and Silverman, 1997)

A systematic review of HL in children with OME by Cai and McPherson (2017) showed that the audiometric configurations of OME related CHL varied between studies, concerning which frequencies are affected more, but generally the low frequencies are affected in all the configurations, with the least affected frequency appearing to be the 2 kHz, followed by the 4 and

8 kHz. Some studies showed that the high-frequency thresholds could sometimes be elevated, especially if there was a long-standing OME (Williamson and Sheridan, 1994; Cai *et al.*, 2017).

Another possible explanation as to why CHL can affect both the low and high frequencies with a lesser effect on the mid frequencies was suggested by Ravicz *et al*. (2004). They studied the effect of introducing normal saline to the middle ear of temporal bone preparations to fill part or all of the cavity. Umbo velocity, which is a direct measure of the dynamic state of the middle-ear ossicles, was measured with a laser vibrometer while the fluid filled all or part of the middle ear. At low frequencies, a reduction in velocity of up to 25 dB depended on the percentage of air left in the middle ear, suggesting that the mechanism of HL at low frequencies is a reduction of admittance of the middle ear space due to displacement of air with fluid. At high frequencies a reduction in velocity of up to 35 dB depended on how much the tympanic membrane was in contact with the fluid, suggesting that high frequencies are affected by the increase in tympanic membrane mass. The latter statement is supported by Silman and Silverman (1997), that a "sloping CHL" reflects a mass tilt, e.g., thickened tympanic membrane (Eapen *et al.*, 2008).

2.2.3.2 Physiology of audibility

To decide which test to apply to assess hearing, it is important to understand the physiology of audibility and how hearing loss affects the outcome of certain measures. A study by Carhart and Tillman (1970) (cited in Plomp (1978)) applied the Northwestern University (NU) speech test to four groups of people with different hearing statuses: normal hearing (NH), CHL, SNHL with good discrimination and SNHL with fair discrimination. The test was applied in quiet (SiQ) and in noise (SiN) with different levels of speech noise ratios (SNRs). The discrimination of the group with CHL was similar to that of NH in all backgrounds. Based on these results, Plomp (1978) proposed that the framework of HL consists of a lack of audibility, which could be measured through PTA and SiQ, and a distortion factor on top of the audibility factor, which can be measured using SiN (Figure 2.6). Later studies showed similar findings (Pekkarinen *et al.,* 1990; Kollmeier *et al.*, 2016).

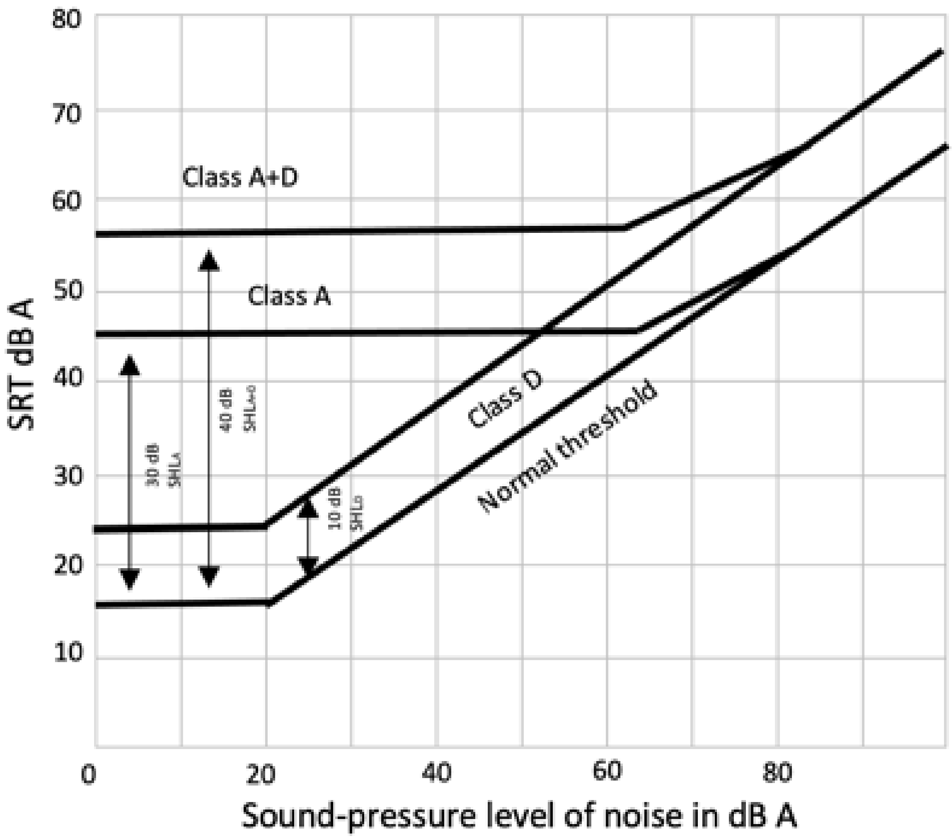


Figure 2.6 Speech perception in normal hearing, in quiet and in noise in Class A, D, and Class A+D (adapted from (Plomp, 1978, Kollmeier *et al.*, 2016): Class A (comparable with attenuation (SHL_A)) involves attenuation of the speech level reaching the inner ear, such as in CHL. Class D (SHL_D) is comparable with distortion, which is present in the case of inability of understanding speech, but no attenuation is present. Cases of (SHL_{A+D}) involve SNHL that present with the two components of HL, attenuation, and distortion, as well as increased signal intensity, requiring improved SNR for the speech to be heard and understood.

Based on this framework, children with OME can be tested for their hearing through behavioural tests such as PTA, play audiometry, and speech tests in quiet such as AMTT (Brown *et al.*, 2022) or objective testing such as ABR, with emphasis on assessment of AC and BC thresholds (Cai and McPherson, 2017).

In their review, Cai and McPherson (2017) showed that articles that reported speech recognition tests in children with OME, including those that used MTT, varied in that most of the speech tests were done in quiet, and a few were done in noise. Details of the studies on SiQ using MTT/AMTT in children with OME are discussed later in the report in section 2.2.4.1.

The decision to measure speech recognition in quiet or noise depends on what is intended to be assessed. Speech recognition in quiet assesses audibility, which is decreased in CHL caused by OME, and it is correlated with PTA hearing thresholds (Hall *et al*., 2007), whereas speech recognition in noise (SiN) assesses both audibility and distortion, which are affected by SNHL. One could argue that children live in a noisy environment, e.g., in schools, day-cares, and playgrounds, and that assessing their ability to recognise speech in noise can provide information on how they perceive speech in such environments. Several studies assessed SiN in children with OME, and the rationale in most studies was that SiN could provide the "true" speech recognition in children with OME considering the environment they live in (Nilsson *et al.,* 1994; Williamson and Sheridan, 1994; Williamson *et al.*, 1997; Lauritsen *et al.*, 2016). One of these studies suggested that SiN testing using the AMTT has an advantage over PTA, with the latter requiring a low ambient noise not exceeding 35 dB (A) (BSA, 2018), a constraint not required for SiN, allowing the assessment of hearing in a noisy clinic (Haggard, 2004). The study by Haggard (2004) showed a correlation between SiN AMTT and PTA hearing levels (PTA-HL) ($r = 0.49$), but the chief interest was the main treatment effect of those measures. The main treatment effect for both the SiN AMTT and HL was significant ($p = 0.004$ and $p = 0.003$, respectively), but there was a significant by-treatment baseline interaction with SiN AMTT baseline, not found with HL. Their study suggested that SiN AMTT can predict the benefit from treatment better than PTA-HL.

Conversely, another article compared reception threshold for sentences (RTS) and sentence recall accuracy (SRA) using the Mandarin HINT in children with actual OME HL and sex and age-matched NH children who performed the tests in normal and simulated CHL conditions (SCHL) (Cai and McPherson, 2017). Their study found that the RTS in noise and quiet were significantly worse in children with actual OME compared to normal children with simulated HL. On the other hand, the SRA in noise at higher SNRs (0-SNR and 5-SNR) was significantly better in children with actual OME compared to those with simulated HL. This result suggested that children with OME HL may develop compensatory strategies to reduce the effects of hearing loss in adverse listening environments (Cai and McPherson, 2017).

Studies have also shown a possible effect of history of recurrent OME at a young age on auditory processing abilities in the long term. A study found that NH children with a history of recurrent OME and ventilation tube surgery before the age of 5 years performed significantly worse in auditory processing tests, including the Gaps in noise (GIN) and dichotic digit test (DDT), compared to NH children with no history of recurrent OME ($p < 0.001$ and $p = 0.002$, respectively) (Khavarghazalani *et al.*, 2022). Their study indicated possible adverse effects of OME on auditory processing when OME is not managed early on.

Considering these points, the mainstay of assessing the hearing thresholds in children with CHL caused by OME is testing the hearing levels using tests such as PTA or play audiometry and/or speech recognition in quiet, especially in young children who cannot perform audiometry. Speech recognition in quiet measures can be used to predict hearing thresholds. However, speech recognition in noise tests may be more sensitive in capturing the everyday listening challenges that children with OME face and the possible long-term effects of OME on hearing. The focus of this PhD thesis is on SiQ.

Further discussion of the studies on MTT can be found in the following section, bearing in mind that in the following sections, the live version of the McCormick toy test is referred to as MTT, and the Automated version is referred to as AMTT.

2.2.4 McCormick Toy Test

2.2.4.1 Overview of the McCormick Toy Test in CHL caused by otitis media with effusion

The McCormick toy test (MTT) is a behavioural hearing test designed and described by McCormick in 1977 (McCormick, 1977). The primary purpose of developing the test was to detect children at risk of hearing loss aged two years and above. The MTT was developed based on the concept and material of previous tests relying on speech recognition to plot speech audiometry for older children, including tests such as the Kendall test, Manchester picture test, and Arthur Boothroyd test (McCormick, 1977). The MTT served the advantage of obtaining information on hearing acuity through assessing their ability to discriminate speech in children ages two years and older due to its simplicity and ease of administration. The test material of the original MTT consisted of seven pairs of actual items (toys) with acoustically similar names (cup/duck, plate/plane, tree/key, spoon/shoe, man/lamb, fork/horse, and cow/house), the two words in each pair had similar vowel diphthong but differing consonants.

The original MTT was performed using live voice by first bringing out the toys individually and encouraging the child to name them. This step was considered essential to ensure the child was familiar with the items and to assess their articulation, noting high frequency consonants and consonant clusters which could be affected by hearing loss. McCormick (1977) suggested that if a child was not familiar with all toys, the test could still be done with fewer pairs, later studies used as few as three pairs of toys (Hall *et al.*, 2007; Lovett *et al.*, 2013), but removal of any pair should be reported in any results given. The examiner could use a sound level meter to adjust their voice level. All the toys were placed on a table in front of the child, the examiner sat one meter away from the child and named the toy at a conversational level, preceded by a leading phrase such as “show me the cup,” and the child would then point at the item. The level of the examiner’s voice would then be lowered to 40 dBA, which is considered the minimal sound level, and if the child identified all the items in front of them at this level, they would have "passed" the test.

A community-based validation study of the MTT using live voice by Harries and Williamson (2000) applied the MTT and standard audiological tests (otoscopy, tympanometry, and PTA) to n = 65 children aged three years coming for their regular health checks. All 14 children who did not pass the standardised tests did not pass the MTT, and the cause was mostly CHL due to OME. Two children were uncooperative and unable to perform the MTT but could perform the standardised test, and one child had a global developmental delay. They concluded that the MTT had a sensitivity of 100% and specificity of 94% in detecting hearing loss, indicating that the MTT, even in live voice, could detect mild CHL (Harries and Williamson, 2000).

Performing the test in live voice had several limitations (Ondáš *et al.*, 2020) including:

- The examiner needed to be well trained to control the level of reproducibility of their voice.
- The difficulty of standardising the test across different testers, testing rooms, and centres.
- Not being able to detect slight hearing loss, such as seen with early OME, which in many instances is characterised by slightly raised low frequencies and normal high frequencies, which could allow for fine consonant discrimination.
- Inability to detect slight unilateral hearing loss.

With the limitations of live voice in mind, automated versions of the Toy Discrimination Test (AMTT) were developed, in which the lead-in phrases and the toy names were recorded, and the stimuli were emitted through loudspeakers or headphones. A study by Ousey et al. (1989) aimed to compare word discrimination threshold SiQ using the IHR-McCormick Automated Toy Discrimination Test, which was a semi-Automated MTT (i.e., the test was computerised, but real toys were used) with those obtained with a live voice. The minimum level for the live voice version was 35 dBA, but the automated version could produce a level as low as 20 dBA. There was a high correlation between the automated and live tests, but the former had better reliability because it was standardised across different examiners and room conditions. The automated test was also more sensitive to mild hearing loss than live voice (Hall *et al*., 2007).

Several studies were conducted to validate the AMTT. A study by Summerfield *et al.*, (1994) was conducted to validate the automated MTT (AMTT) on 187 children, some of which had OME, aged between 2 and 13 years. In their study, children were tested with PTA, tympanometry and AMTT in noise and in quiet. The result of the test was described as the word discrimination threshold (WDT), which is the level at which the child scores 71% (Hall *et al*., 2007). Forty-three percent of the 2-year-olds and 80% of the 3-years-olds managed to complete at least one run of the AMTT. Of the 18 children who did not manage to complete a PTA, eight of them managed to complete at least one run of AMTT, supporting the claim that the Toy Test was easier and more engaging for children to perform, compared to PTA (Summerfield *et al.*, 1994). In their study, Summerfield *et al.*, (1994) also assessed the test-retest reliability of the AMTT, and they found a high correlation ($r = 0.95$) between repeats and small within-subject standard deviation (SD_w), indicating good test-retest reliability. The study by Summerfield *et al.*, (1994) was preceded by another validation study on the AMTT (Palmer *et al.*, 1991), and both studies found a high correlation between the AMTT SRTs and PTA at different frequency range averages (details in section 2.2.4.2).

Hall *et al.* (2007) oversaw an exceptionally large study aimed at identifying developmental changes in word-recognition thresholds (WRT) in children aged two to five years with different middle ear statuses. This was part of the Avon Longitudinal Study of Parents and Children (ALSPAC). Children were tested at 31, 43 and 61 months of age using AMTT (free-field speakers) in an adaptive procedure to identify their 71% WRT in quiet. They adapted the same technique used by Summerfield *et al.* (1994). Tympanometry and PTA were also done. Children who completed all three sessions were n = 762. Children with normal middle ear function showed an improvement in mean WRT by 5 dB as age increased from 31 to 61 months. There was also a significant increase in WRT (worsening) in the presence of OME at 31, 43 and 61 months compared to no OME ($p \leq 0.001$).

A recent study by Brown *et al.* (2019) assessed the ability to discriminate speech in children with OME aged 3 – 6 years using bone conduction (BC) headsets. Brown *et al.* (2019) found a significant improvement in WDT by approximately 15 dB when the children were tested with the AMTT (both in quiet and in noise) using the BC headset, compared to their WDT without wearing the headsets ($p < 0.05$). A later study also investigated the effect of using BC headsets on WDT in NH children using the AMTT, and found an improvement of WDT using BC headsets compared to not using them (Brown *et al.*, 2022).

These studies suggested that the AMTT can reflect changes in middle ear status and that it was correlated well with PTA ($p < 0.001$) (Summerfield *et al.*, 1994; Hall *et al.*, 2007). The AMTT is an easy, engaging test that could be modified as an application on a tablet for clinics where PTA might not be available, potentially improving healthcare delivery for children with OME.

2.2.4.2 Prediction of audiometric threshold from word reception threshold obtained from the McCormick Toy Test

The assumption of correlation between SRTs in quiet and PTA hearing levels (PTA-HL) was based on Plomp's framework of hearing loss which stated that speech recognition in quiet in CHL should be highly correlated with PTA-HL (Plomp, 1978). Several studies analysed the relationship between SRT SiQ and different frequency averages of the PTA-HL (Table 2.3). Note that the term SRT is used instead of WDT (Summerfield *et al*., 1994) and WRT (Hall *et al*., 2007), because it conveys the same meaning as the terms WDT and WRT, and it is the term used through out the PhD.

Table 2.3 The linear regression parameters of the correlation between the SRT (independent variable) and PTA-HL (dependent variable) in different studies (Palmer *et al*., 1991, Summerfield *et al*., 1994; Hall *et al*., 2007)

	Summerfield *et al.* (1994) Palmer et al., (1991)*				**Hall *et al*. (2007)**			
Frequencies (kHz)	**n**	***r***	***β***	**α (dB)**	**n**	***r***	***β* (95%CI)**	**α (dB)**
0.5, 1, 2, 4	84	0.82	0.82	-9.64	962	0.7	0.80 (0.75, 0.84)	-9.47
0.5, 1, 4	98	0.82	0.78	-7.12	962	0.7	0.80 (0.76, 0.85)	-9.56
	66*	0.9*	0.88*	-9.97*				
1, 4	105	0.81	0.84	-10.02	962	0.7	0.82 (0.77, 0.87)	-10.57

(n: sample size, r: correlation coefficient, β: slope, α: intercept)

As seen from previous studies (Palmer *et al*., 1991, Summerfield *et al*., 1994; Hall *et al*., 2007), the AMTT SRTs were strongly correlated with all PTA-HL frequency averages. The following prediction formula was used by the studies mentioned in the table to predict the 0.5, 1, and 4 kHz PTA average of the better ear from the SRT obtained binaurally in a sound field through loudspeakers. This formula was developed using the constant values obtained from parameters of the linear regression of SRT onto average pure-tone thresholds in the better ear:

$$\text{PTA} = \beta \times \text{SRT} + \alpha$$

Where β is the slope of the regression line of the relationship between SRT and PTA, and α was the pure-tone average that would correspond to a SRT of 0 dB(A) (i.e., the intercept of the linear regression). Table 2.4 provides the predicted PTA-HL average of frequencies 0.5, 1, 2 and 4 kHz of the better ear when SRTs are 30 dBA.

Table 2.4 Predicted PTA-HL average of frequencies 0.5, 1, and 4 kHz when SRT71 is 30 dBA (Palmer *et al*., 1991, Summerfield *et al*., 1994; Hall *et al*., 2007)

	SRT (dBA)	**Predicted PTA (dB HL)**
Palmer *et al*., (1991)	30	17
Summerfield *et al.* (1994)	30	15
Hall *et al*. (2007)	30	14.5

The average PTA-HL in the better-hearing-ear can be predicted from the SRT with a 95% confidence limit of ±13 dB or less (Palmer *et al*., 1991, Summerfield *et al*., 1994).

Although there were limitations for using AMTT in measuring hearing, including the fact that only the average thresholds at certain frequencies can be measured, it is still a promising finding for behavioural assessment of hearing in children as young as 2.5 years using a simple and reliable test such as the AMTT or its Arabic version (PAAST).

2.2.5 Developing the Paediatric Arabic Auditory Speech Test (PAAST)

2.2.5.1 Considerations for Testing Children with Speech Tests

When choosing a speech test for children, considerations (listed in Table 2.5) must be addressed to ensure the results are reliable, and that they reflect the area intended to be addressed. Bearing these considerations in mind, the AMTT is a valid, reliable, test that has been equalised for intelligibility (Summerfield *et al.*, 1994; Lovett *et al.*, 2013). It is an easy test that can be performed by a child as young as 2.5 years old who may not be able to perform the conventional audiometry, therefore, it has been chosen to be adapted into the Arabic language.

Table 2.5 Considerations for choosing a speech test for children (adapted from Mendel, 2008; Katz *et al*., 2009)

Considerations	Details
Validity and reliability	Rigorous testing is needed to ensure that the speech test is measuring what it is intended to measure, including ensuring the speech materials are equally intelligible.
Speech test language	The speech test (and the speaker's language if the test was performed in live voice) should be in the child's first language.
Child's chronological age, vocabulary, and language competency	The speech test should include words that are common among children in the same age group of the child, but performance should be independent of the vocabulary knowledge and higher-level abilities.
Developmental auditory skill level	The test should be chosen based on the developmental skill intended to be tested
Type of response task	For example, pointing at a picture or picking up a toy.
Presentation method	Using headphones allows obtaining ear-specific information whereas sound field testing only obtains information about the best-hearing ear (Katz *et al*., 2009).
Presence or absence of reinforcement	The task should be interesting and motivating.

2.2.5.2 Arabic speech tests

There is a limited number of Arabic speech tests in general and automated Arabic speech tests specifically. Table 2.6 lists the available Arabic speech tests for children. There are no validated automated Arabic speech tests for children younger than 5 years, which is around the time OME occurs more frequently (Rosenfeld *et al.*, 2022). Therefore, the need to develop an Arabic automated speech test for children aged 2.5 and above was acknowledged, leading to the development of the PAAST.

Table 2.6 Available Arabic speech tests

Test Name	Author	Stimulus	Age	Presentation of Stimulus	Back-ground	Outcome Measure
Saudi Arabic Speech Audiometry for Adults and Children	Ashoor and Prochazka, 1982	Words	NR	Live voice	Quiet	Speech Recognition Threshold (SRT)
The Arabic SPAC (ArSPAC)	Kishon-Rabin and Rosenhouse, 2000	Words	NR	Recordings (not automated)	Quiet	Percent correct
Speech Perception Test for Jordanian Arabic Speaking Children	Abdulhaq, 2006	Words	6-9 years	Recordings (not automated)	Quiet	Word Recognition Score (Percentage)
Arabic Words in Noise test and Arabic Sentences perception in Noise Test	Abdel Rahman, 2018	Words and Sentences	5-12 years	Recordings (not automated)	Noise	Percent correct
Arabic CHEAR auditory perception test (A-CAPT)	Rayes, Al-malky and Vickers, 2021	Words	5 to 11 years	Automated	Noise	Percent correct

(NR = not reported)

2.2.6 Paediatric Arabic Auditory Speech Test (PAAST)

The PAAST is an Arabic version of the AMTT. The PAAST in noise (PAAST SiN) was developed and validated by Dr. Rania Al-Kahtani (Al-Kahtani, 2020) as part of her PhD thesis at the University of Southampton. The PAAST SiN was designed in the form of an application that can be installed on an iPad, and the child can be tested using calibrated headphones. The test can be done in both quiet and in noise. To understand the scope and limitation of the PAAST, it was important to validate it against other tests, including PTA, and other Arabic speech tests, if available.

The process of developing the PAAST was similar to the MTT (Al-Kahtani, 2020). The word pairs chosen for the test were monosyllabic consonant-vowel-consonant (CVC) words that were acoustically similar, where each pair of words shared the same vowel. Al-Kahtani (2020) assessed the familiarity of the words by consulting audiologists from different regions in Saudi Arabia as well as audiologists from neighbouring Arab countries including Kuwait, United Arab Emirates, Jordan, Palestine, and Egypt. They carefully reviewed the words in terms of applicability to children in their region/country. Assessing familiarity across this broad area was necessary because although the formal Arabic language is standardised across those areas, the non-formal Arabic dialect is different between Arab countries as well as regions within Saudi Arabia. This difference in dialect can extend to difference in terms used to call certain items. The speech material chosen for the PAAST were of words that are similar across countries, but ensuring their familiarity to children from different Arab areas allowed for wider regional use of the tool (Al-Kahtani, 2020). An important aspect regarding the chosen words was that they are phonetically balanced with various phonemes occurring at approximately the same frequency in the Arabic language. The word pairs of the PAAST are listed in Table 2.7.

Table 2.7 The words in PAAST, their phonetic transcription and their meaning in English

Pair	**Word 1 in Arabic**	**Phonetic Transcription**	**Meaning in English**	**Word 2 in Arabic**	**Phonetic Transcription**	**Meaning in English**
1	باب	/ bab/	Door	ناس	/ nas/	People
2	فيل	/ fil/	Elephant	ديك	/ dik/	Hen
3	بيت	/ bet/	House	عين	/ ʕen/	Eye
4	ثوب	/ θɔb/	Mandress	موز	/ mɔz/	Banana
5	نور	/ nuɽ/	Light	دود	/ dud/	Worms
6	دب	/ dub/	Bear	رز	/ ɽuz/	Rice
7	ورد	/ wʌɽd/	Flower	كلب	/ kʌlb/	Dog

The words were recorded by the researcher who developed the PAAST (Al-Kahtani, 2020), a female native Arabic speaker. The speech was recorded in a “white dialect” not related to any specific region in the Arab world. Recordings took place in an anechoic chamber at the Institute of Sound and Vibration at the University of Southampton. A schematic representation of the PAAST SiQ is illustrated in Figure 2.7.

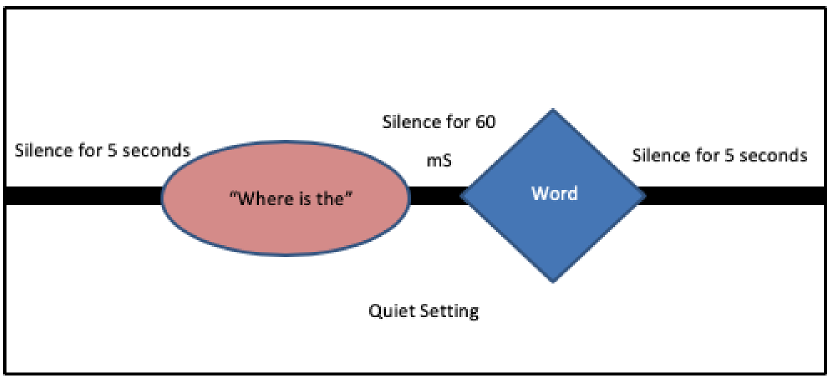


Figure 2.7 Schematic presentation of the PAAST in quiet

Equalisation of intelligibility of the words of PAAST SiN was achieved using the method of constant stimuli (MoCS) (Al-Kahtani, 2020). This was an important step because equal loudness of words does not necessarily indicate equal intelligibility, as some words may consist of sounds with more energy compared to other words, rendering them easier to recognise (more intelligible) (Gelfand, 2016). The PAAST SiN also demonstrated good test-retest reliability (Al-Kahtani, 2020).

Based on the original AMTT, the PAAST was developed to provide a tool that was easy and readily available in the form of an application on iPad to assess the hearing of children as young as 2.5 years.

To apply the test in children to evaluate their hearing, several experiments were conducted in the current PhD to ensure that, in normal hearing adults, the speech material was equally intelligible (Aim 1) and to test the ability of the PAAST and its speech material to detect OME-related SCHL (Aim 2).

2.2.7 Effect of OME on speech perception and language acquisition

OME can lead to CHL, which may involve not only low frequencies, but also, less commonly, higher frequencies (Cai and McPherson, 2017). This hearing loss is expected to have an impact on speech and language development, with secondary communication and behavioral difficulties.

The complexity of the relationship between OME and these complications depends on various factors, some related to the OME condition and some related to the child and their environment (Hall *et al.*, 2014).

Factors related to OME include the length, duration, and severity of the problem (Jamal *et al.,* 2022). Given the self-resolving and recurrent nature of the condition, children with OME could have fluctuating hearing loss. In a case-control study, 12 children aged 8-10 years with history of recurrent OME had significantly worse scores on the Gap in noise test which assesses temporal processing, compared to the sex and age matched control group (Khavarghazalani *et al.*, 2022).

Persistent OME could also influence speech perception and language acquisition. A case-control study of 19 children with history of persistent OME with mean number of tube insertions of 1.4 times, were compared with 19 sex and age matched children with no history of persistent OME (Klausen *et al.*, 2000). The study found significantly worse scores in articulation and sound discrimination tests in the case group compared to the control group (Klausen *et al.*, 2000).

The OME-related CHL may also have a negative effect on sound localisation and binaural processing. Studies have shown that attenuated/ delayed auditory input due to prolonged OME led to the impairment of temporal resolution sensitivity to short tones in the presence of background noise (Newton, 2009). A study on 60 school children with normal tympanometry were assessed for speech recognition in noise showed that children with history of severe OME were at risk of long-term hearing in noise deficits, whereas children with history of mild OME were not (Koiek *et al.*, 2022). Their study suggested that subcortical binaural processing may be affected by severe OME leading to these long-term effects on speech in noise.

Factors related to the child include age, level of existing vocabulary, cognitive and linguistic perceptual abilities, and degree of support at home (Newton, 2009; Hall *et al.*, 2014).

Age affects the stage of the development in the children's ability to acquire speech. The ability to perceive speech is what allows us to externalise language, thus they are tightly connected and synergistic. Speech perception is a process developed early in the human life and undergoes dramatic changes during the first year of life. Infants are born with abilities to acquire any language with their speech perception abilities (Savithri, 2022). Auditory deprivation at any point during childhood before puberty, and especially in the first years of life can negatively influence language acquisition (Savithri, 2022). Conductive hearing loss in the first and second year of life can alter not only the acquisition of sounds, but also semantic, syntactic, and pragmatic rules of language (Newton, 2009).

The environment the child lives in can contribute to severity of the effects of OME related CHL. A study by Hall *et al.*, (2014) examined the association between the number of episodes of OME and HL (OME/HL score) on IQ, and whether that association was moderated by factors including socioeconomic, child, or home. This study by Hall *et al.*, (2014) was part of the Avon Longitudinal Study of Parents and Children (ALSPAC), where 1155 children were assessed regularly between the ages 8 months and 8 years for their OME status and WRT (OME/HL) using tympanometry and AMTT, respectively. Information about the child's cognitive environment was assessed by the HOME inventory which assessed aspects including the number of toys and books the child owned as well as the mother's level of interaction with her child. Children in the highest (worst) 10% OME/HL scores had lower performance and verbal IQ at the age of 4 compared to those who were unaffected, this effect was diminished by the age of 8 years. Home environment showed a moderating effect on the OME/HL score, where children with high OME/HL and low (worse) HOME scores had lower performance IQ compared to those with high OME/HL and high HOME, and this observation was seen up to the age of 8 years. The finding by Hall *et al.*, (2014) sheds light on the importance of cognitive stimulation at home and how hearing loss can be compound by limited cognitive stimulation, further reducing access to verbal interactions and incidental learning, both of which are essential for perceiving speech and consequently acquiring language

(Savithri, 2022). The brain plasticity plays a role in compensating for the negative effects of hearing loss, however if other risk factors are present such as the low cognitive stimulation at home, this impact of hearing loss may be longer.

A recent review by (Homøe *et al.*, 2020) associated OME with negative effects on auditory processing, language and speech development, school readiness, social competence, psychosocial wellbeing, and sleep.

The effect of OME on hearing and language skills can also extend to indirectly affecting the behavior of the child. The gap in acquired speech is stored in the child's language database, and later on children with history of OME may be unable to decipher incomplete auditory messages and find difficulties in using contextual clues or previous experience to decipher the auditory message (Newton, 2009). This gap could lead the child to ask for the volume of speaker's voice to be raised or settling for what they have been able to hear, even if it was not understood by the child, which could further worsen the ability for language acquisition. Early hearing loss due to OME can lead to poor pronunciation which can lead to difficulties in being understood, consequently limiting communication with others and leading to frustration which could impact the psychological well-being of the child and their QoL (Homøe *et al*., 2020).

Brain plasticity can overcome the influence of the periods of speech deprivation, by ensuring sufficient auditory and visual cues are administered (Savithri, 2022). Therefore, early diagnosis and management of OME including auditory training or the use of hearing aids during episodes of OME can improve the outcome of language acquisition and speech perception in challenging environments (such as noisy classrooms) (Brown *et al.*, 2019; Brown *et al*., 2022).

Understanding the possible short- and long-term effects of OME on children indicates the importance of having a good set of outcome measures. It has been recommended that assessments of hearing and QoL should be included as outcome measures in children with OME (Homøe *et al.*, 2020; Rosenfeld *et al.*, 2022).

The following sections discussed the second outcome tool in the current PhD, which is to research QoL as an outcome measure in children with OME.

2.3 Quality of life in children with OME

Objective outcome measures relating to the management of OME, such as resolution, recurrence and improvement in hearing levels are well known important outcome measures, but the importance of subjective measures should also be considered, because they provide an insight of the impact of the disease that may not be possible to assess with other objective measures

(Chessman *et al*., 2016). Over the last two decades, QoL measures have become increasingly recognised as primary clinical outcome measures in research, including studies on OME in children (Gan *et al.*, 2018).

Health-related quality of life (HR-QoL) is defined as a multifactorial construct covering four dimensions: physical complaints, mental condition, functional impairment in daily life and impairment of interpersonal relationships (Homøe *et al.*, 2020).

Health-related quality of life was also defined by Rosenfeld *et al*. (1997) as 'a subjective outcome that reflects the patient's perception of his or her health status', and in the case of young children with OME, the caregiver could be the one reporting this. Another term that can describe the impact of a disease on the HR-QoL is the functional health status (FHS), which is a term considered to overlap with the HR-QoL (Freyers and Machine, 2016), because it provides a way to summarise the impact of the medical and social aspects of health and to link this summary to a HR-QoL outcome. Measuring HR-QoL can be achieved through measuring disease-specific FHS or a generic QoL measure, but the advantage of using FHS is that it comprises both medical and social aspects, allowing for comparison between people with the same health problem. The terminology used to describe the HR-QoL throughout the PhD was QoL.

According to the latest OME practice guidelines of the American Academy of Otolaryngology, QoL assessment is highly recommended as part of the outcome measures after treatment or as follow-up (Rosenfeld *et al.*, 2022), because it has been shown that OME, recurrent OME and AOM are associated with lower QoL (better) scores after treatment (Brouwer *et al.*, 2005). Although the current practice in Saudi Arabia relies on the guidelines of the American Academy of Otolaryngology, QoL is not assessed due to the lack of Arabic disease specific QoL questionnaire for children with OME.

The OM6 questionnaire is a disease-specific QoL questionnaire, and is the most commonly used questionnaire to assess the QoL in children with OME, AOM and COM. OM6 is a valid, reliable, easily scored questionnaire (Gan *et al.*, 2018). The OM6 was chosen to be translated into Arabic and validated so that its ability to measure the QoL in children with OME in Saudi Arabia could be studied. The rationale behind choosing the OM6 questionnaire, including its psychometric functions and the studies that measured QoL using different QoL tools, are discussed in the next sections to support the choice of the OM6.

2.3.1 Psychometric properties of the questionnaire assessing QoL

Psychometric properties are properties that are evaluated to ensure that a questionnaire (or any test) can be adequately used for discriminative and evaluative purposes in research and clinical practice (Timmerman *et al*., 2007). These psychometric properties are reliability, validity, responsiveness and floor and ceiling effects. Details of the properties, including their definition and quality criteria, are listed in Table 2.8.

Table 2.8 Psychometric properties of a questionnaire (*Adapted from* Timmerman *et al.*, 2007; Al Sayah *et al.*, 2013; Tao, Schulz, Donna B Jeffe *et al.*, 2018*)*

Psychometric Property	**Component**	**Definition**	**Quality Criteria**
Reliability	Internal Consistency	The extent to which items in a (sub)scale are intercorrelated, thus measuring the same construct.	Factor analyses performed on adequate sample size. Suggested sample size is calculated as $n = 7i$ Where n is the sample size and i is the number of items in the questionnaire, but ideally n should be >100. Cronbach's alpha(s) calculated per dimension/factor between 0.70 and 0.95 indicates a reliable questionnaire.
	Test-Retest Reliability	Examines results of repeated trials to make sure that conducting the same procedure yields the same results under the same conditions.	Correlation coefficient >0.70 and assessment of the with-in subject variability.
	Stability	A change in response needs to reflect a change in circumstance rather than any other reason.	
Validity	Content Validity	The extent to which the domain of interest is comprehensively sampled by the items in the questionnaire.	A clear description is provided of the measurement aim, the target population, the concepts that are being measured and the item selection, target population, and those (investigators OR experts) involved in item selection.

Psychometric Property	Component	Definition	Quality Criteria
Validity (Cont.)	Criterion Validity	The extent to which scores on a particular questionnaire relate to a gold standard.	A good rating is ensuring the gold standard is 'gold' and a correlation with gold standard is ≥ 0.7.
	Construct validity	The extent to which scores on a particular questionnaire relate to other measures in a manner that is consistent with theoretically derived hypotheses concerning the concepts that are being measured.	A good rating includes those specific hypotheses were formulated AND at least 75% of the results are in accordance with these hypotheses.
	Discriminative validity	The ability of the tool to discriminate between different conditions of the disease, e.g., mild vs severe or none.	Assessed by receiver operating curves (ROCs) and comparison of area under the curves (AUCs). AUC of 1.0: perfect discrimination AUC of 0.5 = no ability to discriminate. AUC should be at least 0.70.
Responsiveness		The ability of a questionnaire to detect clinically important changes over time.	Calculating the change scores and examining the correlation with the corresponding change yielding significant difference. The Standardized Response Mean (SRM), which is the mean score divided by the standard deviation, should be >0.8.
Floor and Ceiling Effects		The number of respondents who achieved the lowest or highest possible score.	Highest or lowest possible response scores should be ≤15%.

2.3.2 The OM6 Questionnaire: Psychometric properties and translated versions

The OM6 is a disease-specific QoL questionnaire developed by Rosenfeld *et al*. (1997), and it was specifically designed to support clinical testing of children with OME. The OM6 questionnaire consists of six items: physical suffering, HL, speech impairment, emotional distress, activity limitations and caregiver concerns. This is a proxy-questionnaire, meaning that it is filled out by the caregiver. There are seven response options to each item: not present/no problem, hardly a problem at all, somewhat of a problem, moderate problem, quite a bit of a problem, very much of a problem and extreme problem. A mean score of all six items comprises the OM6 Total Score. The highest score would be a score of 7, and the lowest score would be 1 (Timmerman *et al.*, 2007). Although there is no normative data set for the OM6, high scores generally indicate a poor

QoL (Gan *et al*., 2018). According to Rosenfeld *et al*. (1997), an improvement of score by <0.5 is considered trivial, a change ranging between 0.5-0.9 is considered small, 1.0-1.4 is considered moderate, and >1.5 change is considered a large improvement.

In general, the OM6 is easy to use, it stands as an overall satisfactory measure of QoL in children with OME (Tao *et al.*, 2018), and it also correlates well with global QoL assessed with generic non-disease specific questionnaires (Kubba *et al*., 2004). According to a review by Maile *et al*., (2013), the OM6 is considered the best measure of FHS in children with OME because of its high reliability, construct validity and responsiveness.

Several studies aimed to validate the OM6 in English and in other languages. Some of these studies were chosen to illustrate detailed evaluation of the psychometric functions (Table 2.9).

Table 2.9 The OM6 validation studies

Aspect	**(Rosenfeld, Goldsmith and Balzano, 1997)**	**(Tao, Schulz, Donna B Jeffe *et al*., 2018)**	**(C.H. Heidemann *et al*., 2013)**
Language	English	English	Danish
Sample size	86	1,045	435
Underlying constructs (factors)		Behaviour and Symptoms (BaS) and Hearing/Speech (HaS)	
Content validity	Good	Good	Not tested
Construct validity	Significant correlation between items and symptoms (*r* = 0.36) and physician visits (*r* = 0.47) (p < 0.001) *Good construct validity.	The significance of correlation between items: Items and BaS: <0.001 Items and HaS: <0.001 BaS and HaS: <0.001 *Good construct validity.	*Good construct validity based on 87.5% of hypothesised correlations.
Criterion validity	Correlation between the score and the global QoL (P<.001). *Good criterion validity	Correlation with the PedsQL (a paediatric global QoL questionnaire) with total score, with BaS and Has: P <0.001. *Good criterion validity	Significant correlation with a gold standard, namely global perceived effect (GPE). *Good criterion validity

Aspect	**(Rosenfeld, Goldsmith and Balzano, 1997)**	**(Tao, Schulz, Donna B Jeffe *et al.*, 2018)**	**(C.H. Heidemann *et al.*, 2013)**
Discriminative validity	Not tested	AUC of OME vs no OME **OM6 total =** 0.840 **BaS items =** 0.867 **HaS items =** 0.621 *Good discriminative ability between OME and no OME, small to no discriminative ability between levels of severity.	Significant difference between rAOM and OME scores <0.001. *Good discriminative ability between rAOM and OME.
Internal consistency	Correlation between items (inter-item) and items and total (items-total) was significant (p-value <0.001). *Good internal consistency	**Cronbach's a value** **Item and item**: >0.7 **Item and total**: >0.7 **BaS items**: >0.8 **HaS items**: >0.8 *Good internal consistency	**Cronbach's a value** **Items and item**: >0.7 **Item and total**: >0.7 *Good internal consistency
Test-retest reliability	Correlation coefficient R > 0.7. *Good test-retest reliability	Not tested	Mean difference in scores between test at day 1 and at day 2 is close to 0. *Good test-retest reliability
Responsiveness	SRM after tube insertion for all items as well as the total score >0.8. *Good responsiveness		
Floor and ceiling effects	*No floor/ceiling effects in total score		*No floor/ceiling effects

AUC: Area under the curve, an AUC of 1.00 indicates perfect discrimination, whereas an AUC of 0.50 suggests that discrimination is no better than chance, SRM: standardized response mean, defined as the mean changed score divided by SD

Overall, the OM6 has a good content and construct validity, although in one study the hearing item was not significantly correlated with the hearing threshold levels of the better ear (Rosenfeld *et al*., 1997). A recent study showed excellent construct validity among all items (Tao, *et al.*, 2018). The criterion validity was good as well (Rosenfeld *et al*., 1997), even though it was not clearly justified why these authors used different tools as 'gold standard' to validate the OM6 against. The discriminative validity of the OM6 was problematic when discriminating between the levels of severity of the disease (Kubba *et al*., 2004; Tao *et al.*, 2018), but it showed good discriminative ability between presence and absence of disease (Rosenfeld *et al*., 1997; Tao *et al.*, 2018) as well as discrimination between AOM and OME (Christian *et al.*, 2013). All studies showed good reliability of the OM6 with good internal consistency and test-retest reliability.

The OM6 can detect response to clinical change (Rosenfeld *et al*., 1997; Timmerman *et al.*, 2007), this was also observed in the randomized clinical trials (RCTs) mentioned in the next section.

In a study by Gan *et al.*, (2018) reviewing the QoL questionnaire in children with OME, they rated the OM6 as one of the top questionnaires in terms of ease in scoring and psychometric properties. One Issue with the OM6 was that it did not convey the HL impact as much as the questionnaires that were designed for this purpose of assessing children with listening difficulties such as Evaluation of Children's Listening and Processing Skills (ECLiPS) questionnaire. Despite that, the OM6 was still considered the tool of choice to adapt in this PhD because it conveyed the FHS of a child with OME and assessed subjective aspects such as the child's behaviour and the caregiver's concern, while hearing can be assessed through measures such as PTA and speech recognition.

The OM6 was also selected to be translated to Arabic because it was the most translated questionnaire for assessing QoL in children with OME. Other reasons for choosing the OM6 were because of its overall good psychometric properties (including content validity, reliability and responsiveness) compared to the other questionnaires, its common use in research, and its ease of scoring and questions (Lameiras *et al*., 2017). Several studies translated (Table 2.10), cross-culturally adapted and tested the psychometric properties of the translated OM6, including the Danish (Heidemann *et al.*, 2013), Portuguese (Lameiras *et al*., 2017) and Dutch (Brouwer *et al*., 2005) versions. Other versions, such as the Malaysian and Turkish, were only translated but not tested for their psychometric properties.

Table 2.10 List of the translated OM6 questionnaires and the validation of the questionnaires in each language

	Danish (C.H. Heidemann *et al.*, 2013)	**Portuguese** (Lameiras *et al.*, 2017)	**Dutch** (Brouwer *et al.*, 2005)	**Malaysian** (Crawford *et al.*, 2017)	**Turkish** (Yazici and Coskun, 2018)
Subjects	435 (OME)	216 patients and parents	384 with recurrent AOM	Not available	Not available
Forward-backward translation	✓	✓	✓	✓	✓
Cross-cultural adaptation	✓	✓			
Content validity	✓	✓	✓		
Internal consistency	✓	✓	✓		
Construct validity	✓	✓			
Discriminant validity			✓		
Criterion validity	✓				
Reproducibility	✓	✓	✓		
Responsiveness	✓	✓	✓		
Floor and ceiling effects	✓				

2.3.3 Studies on the QoL in children with otitis media with effusion

Health-related quality of life has been increasingly acknowledged as an important outcome measure in studies to assess the effect of OME on children. Various databases – including Cochrane Library, Pubmed, Delphis, Saudi Digital Library (SDL) and Google scholar – were searched to review studies that aimed to research QoL as an outcome measure of OME, rAOM and/or COM in children. Studies included for this review were not limited by a year of publication nor by the level of evidence. A total of 19 studies were reviewed, including six RCTs, two metanalyses, and the rest were prospective cross-sectional observational or cohort studies. A table of the reviewed studies that assessed the QoL in children with OME is available in Appendix B. Thirteen studies used the OM6 as a tool to assess the Qol, including four RCTs. Of the 13 studies, six used the English version of the OM6 (Rosenfeld *et al.*, 2000; Chow, Wabnitz and Ling, 2007; Ambrosio and Brigger, 2014; Blank *et al.*, 2014; Els and Olwoch, 2018). One study used the

English OM6 on a Setswana-speaking population with the help of a translator (Els and Olwoch, 2018). Other language versions of the OM6 were applied, including Dutch (Brouwer *et al.*, 2005; Van Dongen *et al.*, 2014), Finnish (Kujala *et al.*, 2014), Greek (Vlastos *et al.*, 2009, 2011), Turkish (Yazici and Coskun, 2018) and Danish (Indius *et al.*, 2018).

The OM6 in a number of studies was used to assess the effect of interventions on QoL, as well as other outcome measures, including the number of episodes and hearing loss. A study assessed the long term effects of grommet tube insertion on QoL, using the Finnish version of the OM6 and found no significant difference between intervention and active monitoring in improving the QoL over the long run (Kujala *et al*. 2014). Another study investigated the difference between two types of interventions: tympanostomy tube and simple myringotomy, and found that both hearing levels and QoL (assessed using OM6) improved similarly with both interventions (Vlastos *et al*. 2011). A study on the efficacy of pneumococcal vaccine found no significant difference in the number of rAOM episodes or QoL (using the Dutch OM6) between the group of children who were given the vaccine (n = 190) and control group (n=193) (Brouwer *et al.*, 2005). These studies suggest that QoL assessed with OM6 can provide a measure of change (or no change) to interventions.

Other studies used the OM6 for conditions other than rAOM and OME. Van Dongen *et al*. (2014) used the Dutch version of the OM6 to study the QoL in children complaining of otorrhea post-surgery who were randomly assigned to three groups: antibiotic-hydrocortisone ear drops, oral antibiotics, and no treatment. The OM6 was applied at baseline and at two weeks of follow-up. Although the OM6 was not usually used to assess the QoL of children with this complication, it was used in this study because it is a validated disease specific QoL Dutch questionnaire, and it assessed aspects similar, but more severe, to those that can be affected in OME. The results showed that the QoL in children treated with oral antibiotics versus those treated with antibiotic drops differed in the median OM6 scores by -2 in favour of antibiotic eardrops, where a change of >1.5 in the mean was considered significant.

A comparison between the OM6 and the questionnaires used in the RCTs, namely the OMO-22 and OM8-30, is illustrated in Table 2.11. The OMO-22 and OM8-30 both had an advantage over OM6 in discriminating severity of the disease, but the OM6 was superior in detecting clinical change. Additionally, the psychometric properties of the OM6 have been extensively tested and proven to be better than those of the other two questionnaires. The fact that the OM6 is the most commonly used tool in RCTs to measure QoL and that it has better properties compared to other questionnaires (Perdrizet *et al.*, 2022) supported the choice of OM6 for adaptation to the Arabic language.

Table 2.11 Comparison between the OM6 and questionnaires that assess HR-QoL used in research

	OM6	**OMO-22**	**OM8-30**
Type	**FHS**	**FHS**	**FHS**
Domains	Physical Suffering, HL, Speech delay, Emotional distress, Limitation of activity, and Caregiver concern	Physical symptoms, Emotional effects, Speech symptoms, Social effects, and Hearing and vestibular symptoms	**Physical health:** Global health, Ear infections, Respiratory symptoms, Sleep patterns **Developmental impact:** Behaviour, Speech and language, School progress, and Parent quality of life
Target age group	6 months - 12 years	6 months - 12 years	3-8 years
Respondents	Parents/caregivers	Parent	Parent
Number of Questions	6	22	32
Scoring	7-point categorial scale	7-point categorial scale	8-point categorial scale
Ease of scoring	Moderate	Easy	Difficult
Validation and Reliability			
Content validity	+	+	+
Criterion validity	?	0	+
Construct validity	+	0	+
Internal consistency	+	+	+
Test-retest reliability	+	+	+
Responsiveness	+	?	0
Floor and ceiling effect	-	0	0
Interpretability	+	+	0

+, positive rating; ?, indeterminate rating; -, negative rating, 0, no information available. Adapted from (Brouwer *et al.*, 2007; Timmerman *et al.*, 2007)

2.4 Gaps in knowledge and Research aims

Otitis media with effusion is one of the most common diseases in children, which could lead to CHL, among other complications, including speech and language delay resulting in behavioural issues and affecting general well-being and QoL (Perdrizet *et al.*, 2022). Recent studies have shown that hearing assessment and QoL measurement can provide information necessary for optimising the management of children with OME clinically and in research (Homøe *et al.*, 2020).

The Paediatric Arabic Automated Speech Test (PAAST) was developed by Dr. Rania AL-Kahtani as part of her PhD project, where the PAAST SiN was assessed for its ability to assess SNHL. The current PhD research assessed the PAAST SiQ as a tool to measure CHL.

The AOM6 is an Arabic disease-specific QoL questionnaire for Arabic children with OME, which was developed in the current PhD. The AOM6 was adapted from the English OM6 questionnaire.

This PhD project aimed to investigate the use of the PAAST and AOM6 as outcome measure tools to assess hearing loss and QoL in Arabic children with OME.

The identified gaps in the knowledge are:

1- Lack of Arabic speech tests for children in general, especially automated tests.
2- Absence of validation studies for the PAAST in quiet, and the lack of knowledge of existing differences in the PAAST validation between SiN and SiQ.
3- Uncertainty regarding how the speech material of the PAAST SiQ in quiet is affected by the audiogram configuration of OME related CHL.
4- Absence of information regarding correlation of PAAST SiQ with PTA.
5- Lack of published validated measures of QoL for children with OME in the Arabic language.
6- Lack of studies on the QoL in Arabic-speaking children with OME. It is unknown if the AOM6 can assess the QoL in the same way the original OM6 does.
7- It is unknown whether there is a correlation between the QoL and speech recognition in children with OME, given speech recognition provides information about how the child is perceiving speech, which could consequently be related to QoL.

2.4.1 Research question and framework of the thesis

The research question in this PhD is: "Can speech recognition and Quality of life measured by the PAAST SiQ and AOM6, respectively, be considered good outcome measures in Arabic-speaking children with OME?".

This question was approached through the following aims addressed in the current PhD:

Aim 1 Equalisation of the intelligibility of the PAAST SiQ speech material.

Aim 2 Ensuring PAAST SiQ sensitivity to OME related simulated CHL.

Aim 3 Developing an Arabic questionnaire to assess QoL in children with OME.

Aim 4 Assessing the use of PAAST and Arabic OM6 in Speech recognition and QoL, respectively, in Arabic children with OME

The experiments that addressed the main aims are listed in Figure 2.8, which illustrates the framework of the current PhD.

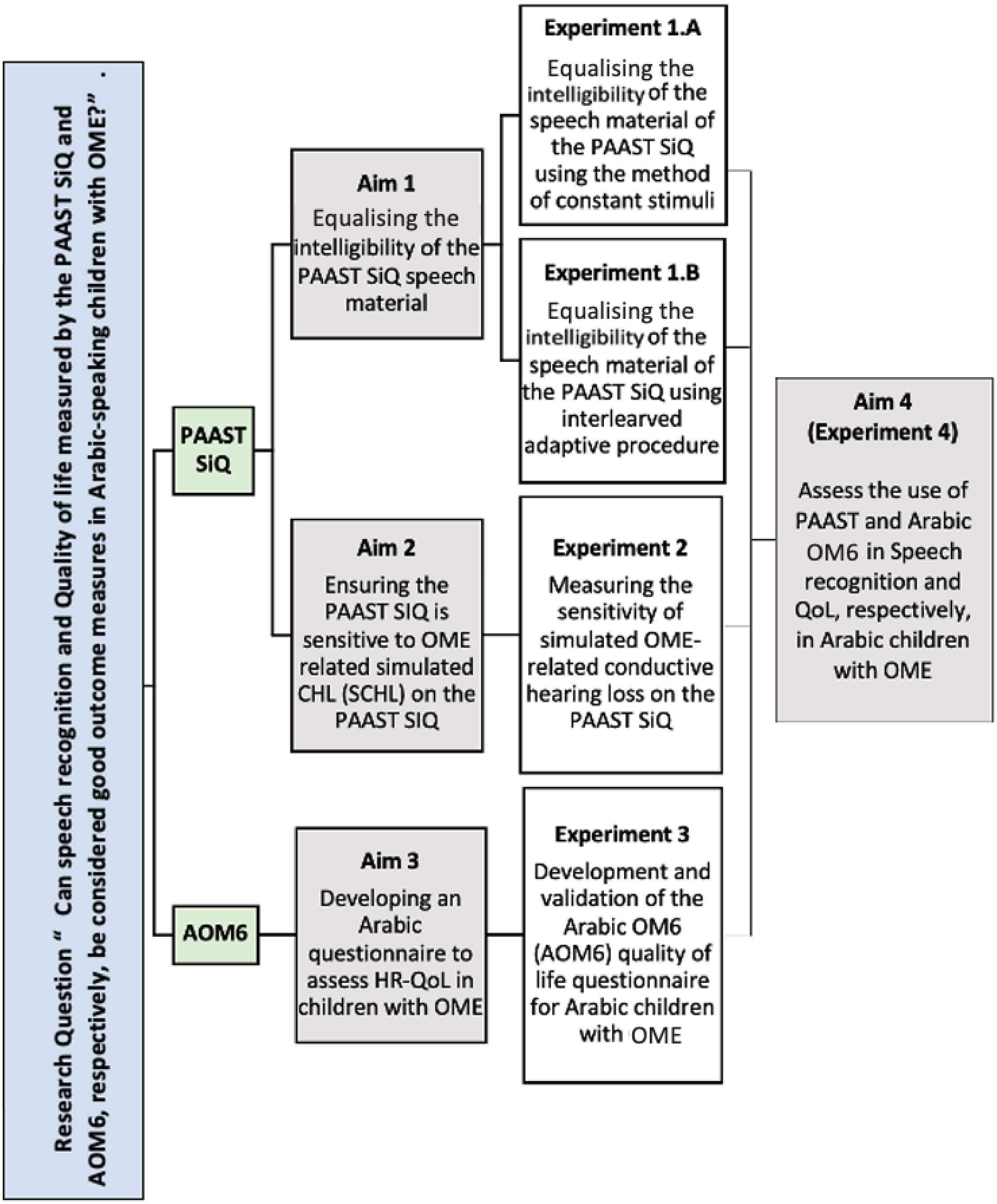


Figure 2.8 Framework of the current PhD project

2.4.2 Aims and summary of the work done

<u>**Aim 1 Equalisation of intelligibility of speech material for the PAAST SiQ test**</u>

In practice, the need for Aim 1 was not obvious from the existing literature when starting the experimental work, so it did not appear as an aim initially. The PAAST was equalised for intelligibility in noise by Al-Kahtani (2020). Based on the literature, tests such as the hearing in noise test (HINT) were equalised for intelligibility in noise only, and these equalised words were used for the HINT in quiet (Nilsson *et al.*, 1994). Therefore, the first experiment (Pilot) attempted to tackle Aim 2 to test the sensitivity of the PAAST SiQ to SCHL. However, the results of the Pilot revealed that the words that had been equalised in noise are not equal in intelligibility when presented in quiet. Therefore, **Experiment 1.A** was conducted to equalise the words of the PAAST in quiet. The view was taken that this was important to address before proceeding with using the PAAST, therefore, the project plan was revised by inserting Aim 1.

Conventional methods were used in Experiment 1.A, namely the method of constant stimuli (MoCS), to obtain the psychometric functions (PF) of the words to equalise the intelligibility of the test. It has emerged, following the careful analysis of the results that MoCS is more limited than previously recognised, at least with obtaining SRTs from the PAAST SiQ, and it has not yet been possible to fully meet the aim. An alternative method, the interleaved adaptive procedure (ILAP), was used to equalise the words of the PAAST SiQ in **Experiment 1.B**. The latter method provided more homogenous results compared to the MoCS, providing a novel contribution in the development of speech intelligibility tests in general.

<u>The broad objectives to achieve this aim were:</u>

1- Equalise the intelligibility of the PAAST SiQ in NH adults
2- Compare the SRTs obtained from applying the test using the MoCS with those obtained using ILAP.
3- Assess the accuracy and precision of the parameters of the adaptive procedure using computerised simulations (Monte Carlo simulations)

<u>**Aim 2 Ensuring the PAAST SIQ is sensitive to OME related simulated CHL (SCHL) on the PAAST SIQ**</u>

As mentioned before, the PAAST SiQ was adapted from the AMTT, and it has been proven that testing with as little as three pairs of words (in case the child was unfamiliar with some of the words) would yield similar results to those who have been tested with all seven pairs (Hall et al., 2007; Lovett et al., 2013). However, it is unknown whether all the words of the PAAST SiQ were sensitive to OME-related CHL or if the elimination of any word pair would affect the sensitivity of the test. Therefore, Experiment 2 assessed the sensitivity of the PAAST SiQ in general and the speech material specifically to OME-related SCHL. This Experiment was conducted by applying the PAAST SiQ on NH adults in an ILAP in three conditions: a normal condition and two conditions of simulated CHL similar to that seen with OME.

The broad objectives of Aim 2 were:

1- Ensure the words of the PAAST SiQ were equally intelligible after the adjustments applied in Aim1.
2- Ensure that the SRTs of all words in the PAAST SiQ were affected by OME-related SCHL.

Aim 3 Developing an Arabic questionnaire to assess QoL in children with OME

This aim was motivated by the lack of Arabic disease-specific QoL questionnaires that assess OME in children and the importance of QoL as an outcome measure. This aim was achieved through Experiment 3 by translating the OM6 from English to Arabic and ensuring cultural acceptance and reliability of the Arabic OM6 (AOM6). The OM6 was not only translated to Arabic but has also been modified to suit Arabic children, and the AOM6 was tested for its psychometric properties. Therefore, it was considered a questionnaire development rather than just a translation.

The broad objectives of this aim were:

1- Translation and cross-cultural adaption of the OM6 to the Arabic language producing the AOM6
2- Assessment of the psychometric properties of the AOM6
3- Assessment of the clarity and cultural acceptance of the AOM6

Aim 4 Assess the use of PAAST and Arabic OM6 in Speech recognition and QoL, respectively, in Arabic children with OME

The previous experiments of the current PhD aimed to ensure that the PAAST SiQ and AOM6 can be made appropriate for assessing speech recognition and QoL, respectively, to continue the process of their validation by applying these measures on children with OME. Aim 4 addressed the main question of the current PhD, through **Experiment 4**. Initially, the plan was to address Aim 4 through a case-control study, where children with OME and as well as NH children would be assessed using the PAAST SiQ. This initial aim was modified to cope with obstacles faced with the COVID-19 pandemic, and the decision was made to include only children with OME who were already coming to the ENT clinic for their appointments and excluding NH children who might face a risk of infection by coming to the hospital for the sole purpose of participating in the research. The decision was also made to assess speech recognition monaurally (each ear) to explore the effect of speech recognition in OME and No OME ears in children with unilateral OME, thus providing some information about the differences in SRTs in the presence and absence of OME.

Aside from exploring the effects of OME status and age on the PAAST SiQ and QoL, the relationship between the PTA hearing levels (PTA-HL) and SRTs was assessed. The relationship between SRTs and QoL scores in OME children was also assessed. Although previous studies investigated the relationship between hearing and QoL, no studies looked at the relationship between speech recognition precisely and QoL. The test-retest reliability of the PAAST SiQ and AOM6 was also assessed to further ensure reliability of these tools

The broad objectives of this aim were:

1- To investigate the effect of OME on speech recognition and QoL.
2- To investigate the relationship between speech recognition and QoL.
3- To assess the reliability of the of the PAAST SiQ and AOM6.

Chapter 3 Experiment 1.A Assessment of the homogeneity of the words in the PAAST SiQ using the method of constant stimuli (MoCS)

3.1 Introduction

This chapter addressed Aim 1, which was to equalise the intelligibility of the words of the PAAST SiQ. The first element in this chapter was the pilot study, originally directed towards testing the sensitivity of the PAAST in quiet to simulated OME related CHL. This was the first experiment conducted in this PhD, but since its results indicated the need to revisit the intelligibility of the words of the PAAST in quiet (Aim 1), the experiment was considered a pilot study. Addressing the sensitivity of the PAAST SiQ to simulated CHL (Aim 2) was achieved in the Experiment 2 (Chapter 5).

The second element in this chapter was to assess homogeneity of the words of the PAAST SiQ using the MoCS. The results of this study served as justification for the additional experiment (Experiment 1.B in Chapter 4) that fully addressed Aim 1.

3.2 Pilot study of the effect of simulated OME-related hearing loss on the PAAST in quiet

Speech recognition using the PAAST in quiet (SiQ) has been proposed in the current PhD as an outcome measure to assess CHL in children with OME. The PAAST SiQ is adapted from the MTT, and was used as a tool to assess speech recognition and hearing in children with OME (Hall *et* al., 2007). In their study, Hall *et* al. (2007), found that testing children with fewer word pairs did not affect their speech recognition thresholds, but the eliminated word pairs were not specified in their study. The current study (pilot) was designed to assess the sensitivity of PAAST SiQ to OME-related SCHL and to determine whether eliminating word pairs would alter the PAAST SiQ results.

3.2.1 Aims and objectives

Aim: Assessing the sensitivity of the PAAST SiQ to OME-related SCHL

The objectives directed towards achieving the aim were:

1. Recruit a sample of Arabic-speaking otologically normal adults.

2. Use SCHL to recreate the low frequency rising audiogram that typically occurs with OME. These simulated audiograms varied in terms of the cut-off frequencies at which the low frequency rising HL loss starts.

3. Obtain the psychometric functions (PFs) of each participant in four hearing conditions: one NH condition i.e., no simulation applied, and SCHL conditions with different cut-off points.

4. Obtain the PFs of each of the 14 words averaged across all participants in four hearing conditions.

5. Compare the PFs of all 14 words in the four conditions.

3.2.2 Method

3.2.2.1 Study design

Data collection from NH adults was carried out in the Audiology clinics at King Abdul-Aziz University Hospital (KAAUH) in Jeddah between December 2018 and March 2019. Participants were tested on their ability to recognise speech at different sound intensities.

All testing was performed in a soundproof room, with a background noise level no greater than 30 dBA.

The experiment lasted 2.75 hours and was composed of three sessions. Session 1, a screening session, which lasted 15 minutes, and sessions 2 and 3, the main testing sessions, which lasted 75 minutes each. Participants attended the sessions on separate days. Participants were given the opportunity to take a five-minute break every 20 minutes during testing.

3.2.2.2 Participants

A total of 20 (16 females and four males) Arabic speaking, otologically healthy participants aged between 18 and 45 years (mean = 28.3 years) were recruited via email and posters advertising the study. This sample size was based on a previous PhD study that examined the validation of the same test in noise. All participants were native Arabic speakers living in Saudi Arabia, they were all Saudi except for one Palestinian and two Yemenis. Inclusion criteria were NH Arabic speaking adults aged 18-45 years. Recruitment of participants was done by advertising the study by a flier displayed on bulletin boards in King Abdulaziz University Hospital. The flier included a description of the study and a Google form link (barcode) that allowed candidates to enter their contact information and fill the otological health form. Eligible candidates were contacted to arrange for their hearing screening. Inclusion criteria and otological health forms are attached in Appendix C.

3.2.2.3 Session 1

Otoscopy was performed for eligible candidates according to the British Society of Audiology's recommendation procedure for ear examination (BSA, 2016) to exclude any abnormalities in the external canal and/or the middle ear. Tympanometry using a probe frequency of 226 Hz was performed according to the BSA's recommended procedures for tympanometry (BSA, 2013). Each participant was then tested with PTA using the standard clinical procedure for PTA in accordance with the BSA (2018) recommended procedures. The participant wore TDH49 supra-aural headphones, heard pure tones of different frequencies and intensities and was instructed to press the response button if they heard them, even if the tone was very low. The PTA was performed at frequency range from 0.25-8 kHz monaurally (each ear) with a maximum starting level of 60 dB HL using an Interacoustics Audiometer (AC40). Normal hearing was defined as AC hearing threshold ≤ 15 dB HL at 0.25-8 kHz frequencies. The researcher conducted the screening.

3.2.2.4 Sessions 2 and 3

The PAAST in quiet was applied through MATLAB software using code specific for this study. The test consisted of a graphical interface containing 14 pictures of acoustically similar Arabic names and a headphone (Sennheiser HD 650) connected to the computer.

The participant faced a monitor with the interface containing the test pictures on display. They were instructed to listen to the sentences ('where is the' + name) through the headphones connected to the computer and click with a mouse on the corresponding name of one of the 14 pictures. They were instructed to guess if they did not hear the word or if they were unsure and were encouraged to choose the closest word to what they may have heard. The participants' responses were automatically recorded on the computer.

The phrases were presented at different intensity levels throughout the test, with a maximum intensity level of 30 dBA.

The method of constant stimuli (MoCS) was used to obtain the psychometric functions (PF) for each word. The MoCS was used by Summerfield *et al.* (1994) to obtain the SRTs of each word to equalise the intelligibility of the speech material. Since the aim was to compare the effect of SCHL on each word, it was decided to choose a method that would allow obtaining the PF parameters for each word, namely the MoCS. Details of how the method of MoCS is administered and fitting a PF are mentioned in (Appendix D.3).

The stimulus levels were determined based on preliminary level measurements during pilot experiments and were chosen to include the full range of the speech intelligibility score (0% to 100%). The stimulus levels chosen for each condition are listed in Appendix E.

In session 2, the participants listened to 14 words repeated three times with a total of 42 words at each of the six levels in the four different conditions. Each level took approximately three minutes to complete, with a total of 75 minutes for the session. Session 3 was the same as session 2 and was performed on a different day. Results from sessions 2 and 3 were then merged for analysis, giving a total of 6 repeats for each word in each condition. The test was divided over two sessions to avoid fatigue. The stimulus levels of each condition were presented in random order with a total of 24 levels for all conditions.

3.2.2.5 Simulated conductive hearing loss

The participants were tested with the PAAST in quiet in four conditions. These conditions were chosen based on the most common audiometric configurations of the hearing loss seen with OME, which is low-frequency hearing loss. All simulated conditions were low-frequency hearing losses varying in the cut-off frequencies after which the thresholds are not attenuated. Threshold elevation was used to simulate hearing loss and was designed using a code in MATLAB. Details of the CHL simulation are mentioned later in section 5.1.1.1, where the experiment was revisited as Experiment 2 (Chapter 5). The four conditions were:

a. Normal hearing
b. SCHL 2K (low frequency rising with cut-off frequency at 2 kHz)
c. SCHL 4K (low frequency rising with cut-off frequency at 4 kHz)
d. SCHL 8K (low frequency rising with cut-off frequency at 8 kHz)

3.2.2.6 Material, calibration, and safety

The procedure was performed using a MacBook laptop with an installed MATLAB application with the code for the PAAST SiQ, Sennheiser HD 650 headphones, and Babyface Pro sound mixer. The laptop and headphones were connected to the sound mixer to process the output from the laptop to the headphones.

Calibration was conducted as described in ISO 389-8 for calibration of headphones for audiometric purposes. Objective calibration of the stimulus was conducted using BRÜEL & KJÆR TYPE 4231 sound level meter and an artificial ear type 4153, using equivalent continuous sound level (L_{eq}) parameter. The stimulus used in the calibration was 1 kHz pure tone of 50 dBA intensity.

The objective calibration took place at the beginning of the experimental period and then every 2 weeks thereafter.

The speech stimulus never exceeded 50 dBA and was primarily presented near absolute hearing threshold of the participants. This was well below the maximum safe level for eight hours of exposure per ear per day as outlined in the ISVR Report 808.

3.2.2.7 Ethical considerations

Ethical approval (ERGO II: 46958) was granted on 24 December 2018 (Appendix I.1). Ethical approval for the application (569-18) to King Abdulaziz University was approved on 13 November 2018 (Appendix I.2).

3.2.3 Analysis strategy

Data were analysed using MATLAB v2018b software through a code written by the research supervisor (Dr. Daniel Rowan). The percent correct identifications of each test word at each level averaged across all 20 participants was calculated.

MATLAB Palamedes Toolbox software was used to generate and fit the logistic regression function to the percent correct identifications of each word across the different levels for each participant and word, and the PFs were plotted against the percent correct (y axis) and intensity level (x axis).

Listed below are the parameters of the PFs obtained. Details of the PFs and methods of obtaining the data are mentioned in Appendix D.

<u>The parameters of the PF:</u>

- **Location:** The location represents the intensity corresponding to the steepest point of the PF. Previous studies (Summerfield *et al.*, 1994; Ozimek *et al.*, 2010; Semeraro *et al.*, 2017) used SRT, the intensity level corresponding to a fixed percentage correct (e.g., 50%), to measure performance in speech tests. Location and SRTs are similar in value. However, location is considered a fixed point on the PF, while SRT is the threshold corresponding to 50% correct. Therefore, the location can be considered a more general approach to assess the homogeneity of the words, allowing the speech material to be utilised in different methods (e.g., adaptive procedures) (Al-Kahtani, 2020). The intelligibility of the words is estimated from the location by finding the difference between the location of the words and the mean location across words and participants (grand mean). In the current study, the words of the PAAST were equalised in noise, so the difference in location from the grand mean in the normal condition was expected to be no more than ±1.5 dBA. A difference outside this range was considered not equal in intelligibility.

- **Slope:** The slope of the PF represents its gradient and indicates the relationship between the change in percent correct identifications and the variation of intensity levels, so the smaller the slope value, the shallower the PF. The slope calculated from the logistic function fitting of the PF has no unit. The slopes of the words were expected to be ≥0.3, similar to those yielded from the post equalisation of the PAAST in noise (Al-Kahtani, 2020). To express the slope value obtained from the PF as percent %/dB, an equation can be applied (Strasburger, 2001):

$$\beta' = ((1-\gamma)/4).\ \beta$$

Where β' is a numerical value that is derived from the slope, which is independent from threshold, γ is the guess rate, which in this study is a value equal to 0.07 (=1/n where n is the number of words, which are 14), and β is the slope.

$$\beta' = ((1-0.07)/4).\ \beta$$

The guess rate (0.07) was rounded to the nearest one, which was equivalent to 0.

$$\beta' = ((1-0)/4).\ \beta$$

$$\beta' = (1/4).\ \beta$$

To get the slope expressed as %/dB:

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

Throughout this thesis, slopes are expressed in the form of β, but the slopes are explained in the form of %/dB in some instances for clarification.

Accuracy of the parameters or functions

- **Standard error (SE) of location:** This represents the distance of the data from the PF. The smaller the SE, the more accurate the data. An SE of ≤3 dB was considered acceptable, an SE ≥3 and ≤8 was considered high, and an SE ≥8 was considered unacceptable and could be labelled as an ERROR.

- **pDev:** Goodness-of-fit, pDev, is a measure of how well the data fit the PF. A value greater than >0.8 was considered very good fit.

3.2.4 Results

3.2.4.1 Ensuring the words of the PAAST in quiet are equal in intelligibility

The logistic regression function was used to fit the data points using MATLAB Palamedes Toolbox software. The percentage of correct identifications of each test word at each level averaged across all 20 participants is plotted in Figure 3.1, and the results are listed in Table E.2, Appendix E.

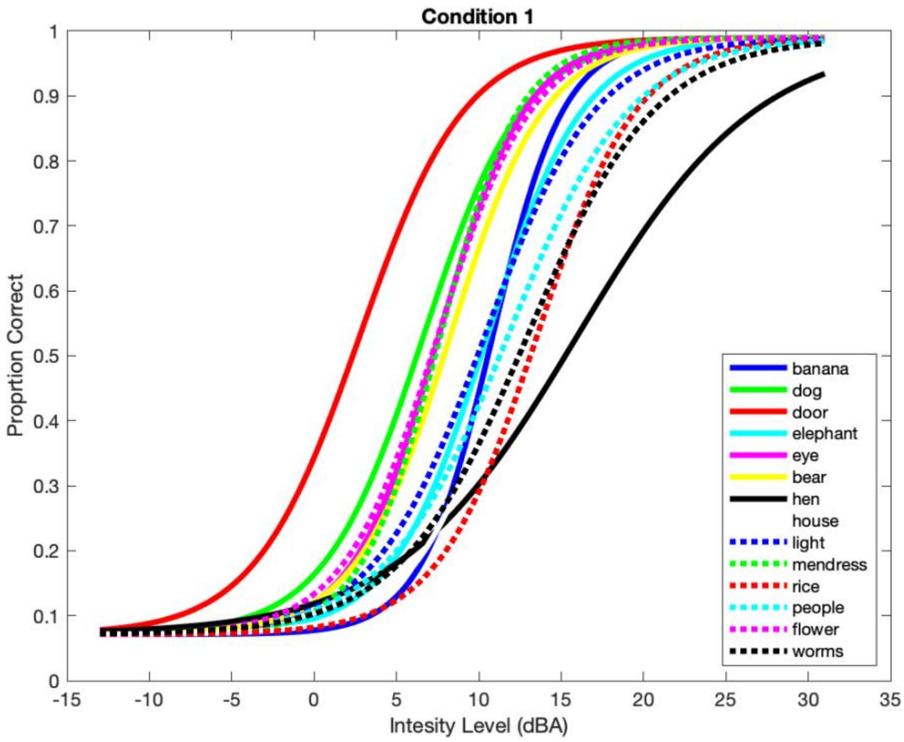


Figure 3.1 The psychometric functions of each word averaged across all participants in the Normal condition

The grand mean location of the words in the normal hearing condition was 9.89 dBA. The word HEN yielded the highest location (15.98 dBA) while the word DOOR yielded the lowest location (2.78 dBA). The difference of locations of the words from the grand mean location ranged between -7.11 and 6.09 dBA, which was higher than the acceptable range of difference of ±1.5 dBA, indicating high variation between locations of the words considering the words of the PAAST were equalised in noise.

The variations between locations of each word among the participants are displayed in a boxplot in Figure 3.2.

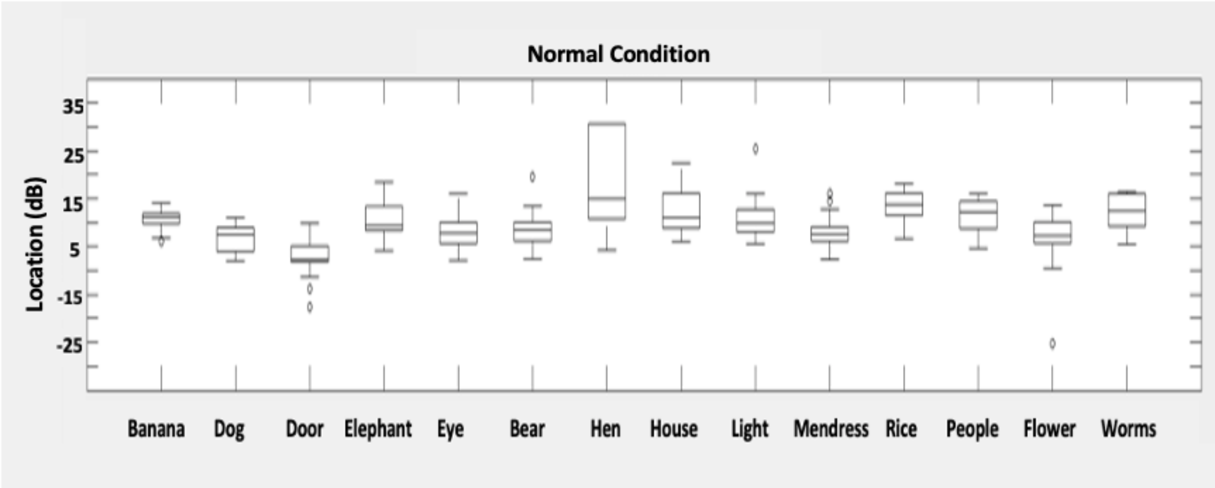


Figure 3.2 Boxplot of the mean locations of each word obtained by each subject (n = 20) in the NH condition; 'normal' refers to testing in the NH condition, i.e. no CHL simulation applied

The slopes of the words ranged between 0.2 and 0.5 (4.6 and 11.5 %/dB). The word HEN showed the shallowest slope (0.2), while the word DOG had the steepest slope (0.5). Most of the words showed a slope ≥0.3, indicating reasonable slopes overall.

All the words showed good pDev > 0.9, which was considered a very good fit.

The SE of the locations of four words (HEN, RICE, PEOPLE and WORMS) were extremely high (>100), representing a large variation in responses to these words among all participants.

3.2.4.2 The effect of the SCHL on the location of the words

The PFs for the three SCHL conditions were obtained in the same way as the NH condition. The main results in this section are directed towards providing preliminary results for Aim 2.

A boxplot (Figure 3.3) of the difference between the mean locations across words for all participants in the four different hearing conditions demonstrated an increase in location with SCHL. These differences confirmed that the chosen audiograms for simulation were reasonable for testing the effect of CHL on the words in the PAAST in quiet.

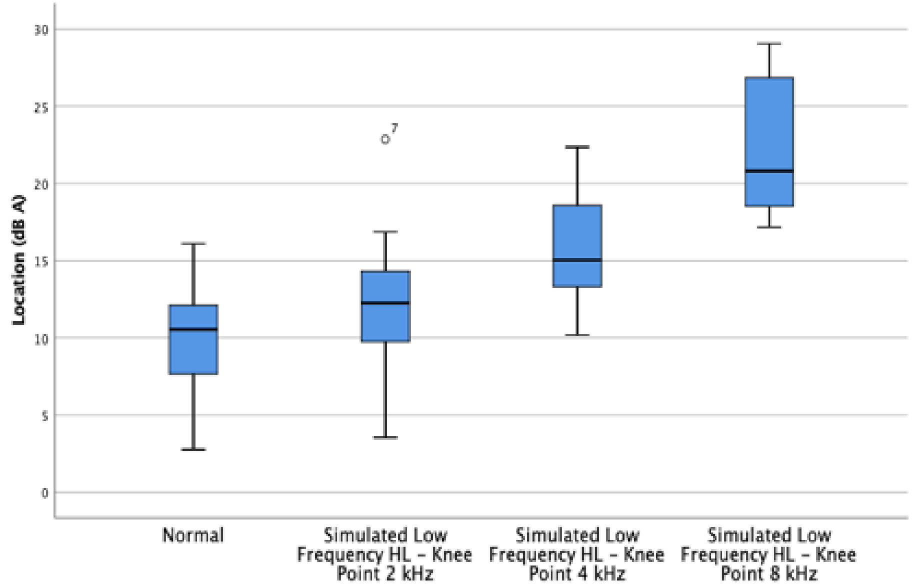


Figure 3.3 Boxplots of the average location across the words in each subject (n = 20) (i.e. 20 points in each boxplot) in the four hearing conditions

No further analysis was discussed regarding the effect of SCHL on the PAAST due to the unequal intelligibility of the words in the NH condition, which rendered the results of the other three conditions unreliable.

3.2.5 Discussion

In the NH condition, the range of locations of the words showed high variation between locations obtained from the PFs of the words across participants. This finding was not anticipated considering the words used for the PAAST SiQ were equalised for intelligibility in noise by a previous researcher (Al-Kahtani, 2020), and some studies equalised speech tests in noise only, and applied them in SiQ (Nilsson *et al.,* 1994; Wong and Soli, 2005). The range of difference in location of the words from the grand mean location obtained from the PAAST in noise (SiN) after equalisation was ±0.8 dB (Al-Kahtani, 2020), which was much smaller than that obtained from the PAAST SiQ (-9.8 – 6.09 dB).

These findings suggested that the words of the PAAST should be equalised for intelligibility in quiet.

Based on the word locations in the NH and simulated conditions, the audiograms of the SCHL were deemed sufficient to represent HL due to the increase in location with increasing cut-off frequency, allowing for using the same conditions in Experiment 2 (Chapter 3).

3.2.6 Conclusion

The words of the PAAST SiQ were not equally intelligible even though the words have been equalised for intelligibly in noise. Therefore, Experiment 1.A was conducted to equalise the words of the PAAST in quiet.

Regarding the effect of CHL simulation on the PAAST in quiet, the audiograms chosen to simulate the HL were reasonable and were used for Experiment 2, which aimed to assess the effect of SCHL on speech recognition using the PAAST SiQ.

3.3 Experiment 1.A: Equalisation of intelligibility of the PAAST in quiet using Method of constant stimuli (MoCS)

3.3.1 Introduction

The results from the pilot study showed that the words of the PAAST were not equal in intelligibility in quiet. The current study discussed the pre-equalisation step of the PAAST SiQ using the MoCS that attempted to address Aim 1.

3.3.1.1 Equalisation of intelligibility of speech stimuli.

In speech recognition tests, equalisation of intelligibility between the stimuli is crucial to ensure that all the words/sentences of a speech test have the same effect on the threshold for all these stimuli, thereby ensuring that the speech test is reliable (Plomp and Mimpen, 1979).

The speech intelligibility of the words of the PAAST SiN was ensured by Al-Kahtani (2020), where they used the MoCS to obtain the locations of the PFs of each word. The equalisation of the words of the PAAST SiN was achieved by adding or subtracting from the root mean square (RMS) amplitude value of each word by the value of deviation from the mean location of the PF (Al-Kahtani, 2020). The MoCS was also used by Summerfield *et al.* (1994) to obtain the SRTs of each word to equalise the AMTT. As mentioned before, using the location of the PF as a reference point ensured adjustment levels that are based on the steepest point in the PF.

The current experiment attempted to equalise the words of the PAAST SiQ by obtaining the locations using the MoCS, by applying more levels than those used in the Normal condition in the Pilot.

3.3.1.2 Equalisation of speech tests in quiet and in noise

Considering the results of the pilot, equalising the speech intelligibility of a speech test in noise alone does not ensure equal speech intelligibly in quiet. A review of studies that aimed to equalise speech test material in noise and quiet are listed in Table E.3 (Appendix E).

Several studies that aimed to equalise speech recognition tests, including the HINT (Nilsson *et al.,* 1994) and the Chinese version of the HINT (Wong and Soli, 2005), equalised the words of the HINT SiN and assumed the words were equalised in SiQ as they were in noise.

An earlier study (Loven and Hawkins, 1983) assessed the inter-list intelligibility of the CID W-22 Auditory test developed by Hirsh in 1952 (Hirsh *et al*., 1952), which was a test composed of word lists that were (according to the developers) equivalent in their average level and range of difficulty. The test was applied to participants in quiet and in multi-talker babble noise. The lists were equal in intelligibility in quiet but were not equally intelligible in noise (Loven and Hawkins, 1983). The AMTT (Summerfield *et al*., 1994) was equalised for intelligibility in noise and in quiet separately, where the SRTs obtained using the MoCS in the pre-equalisation stage had different ranges in quiet compared to those in noise.

A review on the effect of background noise on the slopes of the PF (Macpherson and Akeroyd, 2014) showed that threshold and slope can vary with different background noise, leading to unequal levels of intelligibility between the stimuli.

Similarly, the pilot conducted in the first part of the current study showed the range of locations obtained from the PAAST SiQ were large compared to those seen in the PAAST in noise. One important reason for this was that the SRTs SiQ (or in this case locations) are highly correlated with the PTA thresholds of the participants, unlike testing speech in a noise background (Nilsson *et al*., 1994), which assesses audibility and distortion components of hearing loss.

It was therefore recommended to equalise the intelligibility of the words of the PAAST in quiet before proceeding with any experiment to ensure reliable results. The pre-equalised words of the PAAST SiN were also tested in the current study to compare the results to those obtained by Al-Kahtani (2020), to confirm that the difference in results between SiQ and SiN was related to absence or presence of noise background.

3.3.2 Aims and Objectives

The aims that addressed the research question were:

1- To equalise the intelligibility of the words of the PAAST SiQ

2- To compare the levels of adjustment needed for equalisation of the PAAST SiN in the current experiment with those obtained from the study by Al-Kahtani (2020).

- The objectives to achieve the aims were:

1- Recruit a sample of Arabic-speaking otologically normal adults.

2- Apply the PAAST SiQ using more levels for the MoCS than those used in the pilot study (Normal Condition) and obtain the PFs of each word.

3- Apply the PAAST SiN using the original sound files of the words for SiN (unequalised for intelligibility) and measure the PFs of each word.

4- Obtain the levels of adjustment needed for adjusting the RMS amplitude of the words.

5- Compare the findings of the pre-equalisation session of the PAAST in noise in the current experiment with those from a previous study that equalised the PAAST SiN.

3.3.3 Method

3.3.3.1 Study design

Data collection from NH adults was carried out in the Audiology clinics at KAAUH in Jeddah between December 2018 and March 2019. Participants were tested on their ability to recognise speech at different sound intensities.

All testing was performed in a soundproof room, with a background noise level no greater than 30 dBA.

The experiment lasted approximately 90 minutes and was composed of three sessions. Session 1 lasted 15 minutes. Sessions 2 and 3 lasted 35 minutes each. Participants were given the opportunity to take a five-minute break every 20 minutes during testing.

3.3.3.2 Participants

A total of 20 (12 females and eight males) otologically normal participants aged between 18 and 45 years (Mean = 29.39, standard deviation (SD) = 5.68) were recruited via email and posters advertising the study. Participants were both KAUH students/staff and people from outside the university. The participants were different from those who participated in the pilot study. The decision on the sample size was based according to previous studies that aimed to equalise speech stimuli (Summerfield *et al.*, 1994; Semeraro *et al.*, 2017; Al-Kahtani, 2020) because there was no method to calculate the sample size for such test, and the aim was to collect enough data to obtain accurate estimations of PF parameters and calculate the levels of adjustment. The sample size decided on was n = 20 which was similar to the sample size in the PhD study by Al-Kahtani (2020) that examined the validation of the same test in noise. All participants were native Arabic speakers living in Saudi Arabia. All of them were Saudi, except for two Yeminis. The inclusion criteria and recruitment method were similar to those in Section 3.2.2.2.

3.3.3.3 Session 1

Same as the Section 3.2.2.3.

3.3.3.4 Session 2 and 3

The PAAST SiQ followed the same test procedure as explained in the Pilot experiment in the Normal condition (Section 3.2.2.4), except that more levels were included in applying the MoCS. In addition to the PAAST SiQ, the PAAST SiN was tested using the same procedure as the PAAST SiQ. The masking noise in the PAAST SiN was white noise, modified to match the power spectrum of the PAAST sentences. The masker started 300ms before the speech signal and ended 300ms after the signal. The noise level was kept at a constant 65 dB SPL and the SNR was determined by the sentence intensity level. The parameters set for PAAST SiN were the same used by (Al-Kahtani, 2020). The intensity levels chosen for the PAAST SiQ and SiN are listed in Table E.4 (Appendix E).

3.3.3.5 Analysis strategy

The analysis strategy was similar to the pilot (Section 3.2.3), with the addition of Confidence Intervals (95%CI), in order to estimate whether the locations from the previous studies fall within the 95%CI. The 95%CI was calculated using the following formula:

$$95\%CI = 2 * SE$$

3.3.3.6 Material, calibration, and safety

<u>PAAST in quiet:</u> Same as the pilot section 3.2.2.6.

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

3.3.3.7 Ethical considerations

Ethical approval (ERGO II: 46958.A1) for this experiment was granted on the 16th of April 2019 (Appendix I.3). Ethical approval for the application (569-18) to King Abdulaziz University was also approved on 13 November 2018 (Appendix I.2).

3.3.4 Results

3.3.4.1 The PAAST SiQ

The grand mean location of the words was 12.75 dBA. The lowest location was observed for the word EYE (8.9 dBA). The highest location was found for the word HEN (19.5 dBA). The difference in locations of the word from the mean location ranged between -4 and 7 dB (Figure 3.4).

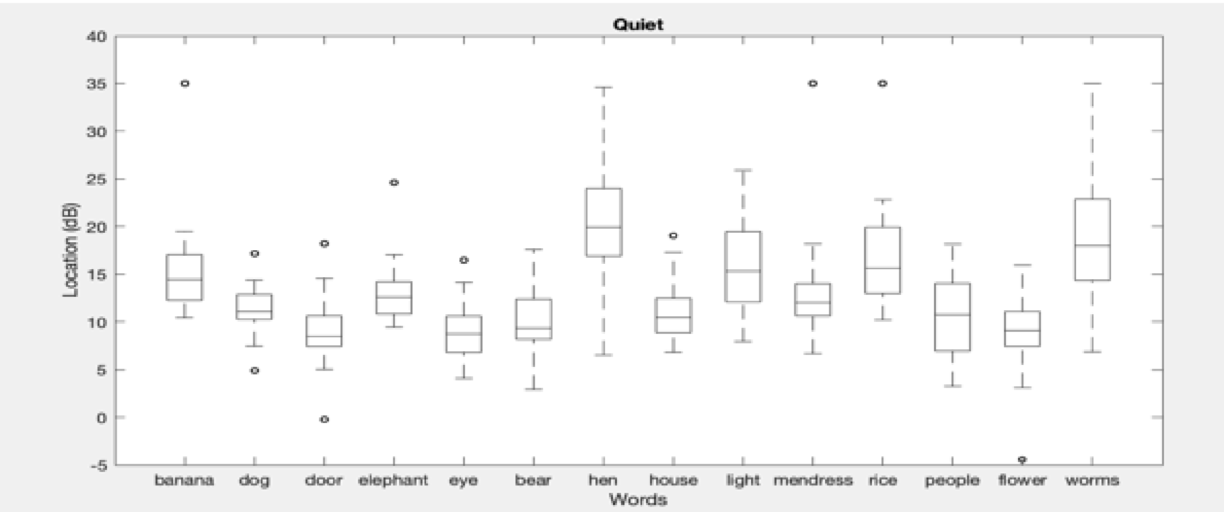


Figure 3.4 Boxplots of the locations of each word across all participants when tested with the PAAST in quiet (each boxplot represented the locations of a word in n=20)

The mean location of the words HEN and Worms (19.47 and 18.09 dB, respectively), were close to the highest intensity presented (20 dB), suggesting that the words were difficult to recognise at the highest level, and higher testing levels would be needed to obtain an accurate location.

The slopes of the words ranged between 0.2 and 0.4 (4.6 and 11.7 %/dB). The word HEN had the shallowest slope 0.2, while the word EYE had the steepest slope 0.4.

3.3.4.2 The PAAST SiN

The locations of the words in the PAAST in noise ranged between -15.9 and -9.6 dB SNR. The grand mean location was -12.9 dB SNR. A boxplot of the locations of each word across all participants plotted in (Figure 3.5). the results of the PF parameters are listed in Table E.6.1 (Appendix E).

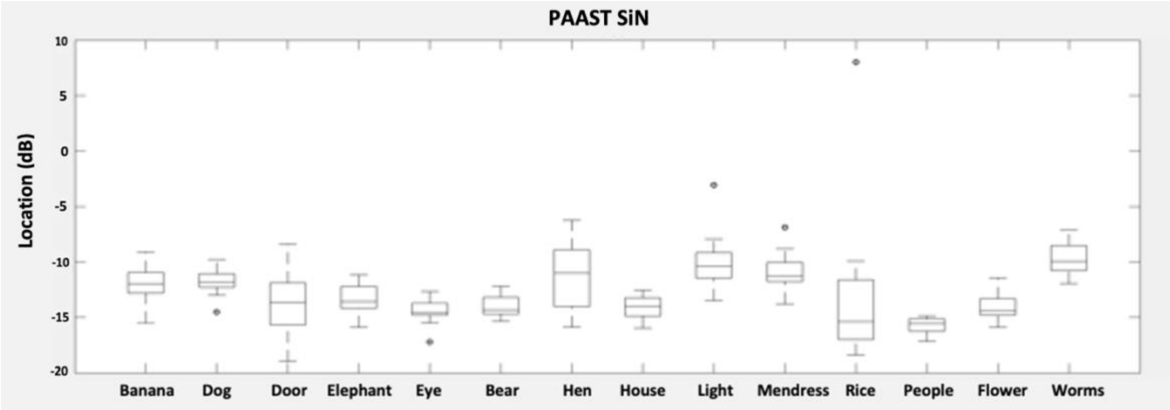


Figure 3.5 Boxplots of the locations of each word across all participants when tested with the PAAST SiN (each boxplot represented the location of the word in n=20)

A comparison between the results of the study by Al-Kahtani (2020) and the current study is summarised in Table 3.1. The PF plots from both experiments are attached in Appendix E.6.

The grand mean location from the study by Al-Kahtani (2020) was close to 95%CI of the of the grand mean location obtained from the current study, indicating the grand mean locations from the two studies were similar. The slopes and the range of differences from the grand mean location from both studies were similar as well.

Table 3.1 Comparison between the parameters of the pre-equalisation stage of the PAAST SiN in the study by Al-Kahtani (2020) and the current study.

	Al-Kahtani (2020)	**Current study**
Sample size	30	20
Grand mean ocation (dB SNR) (SD (dB))	-11.1 (2.0)	-12.9 (1.9)
95%CI (dB) of the grand mean location	-12.2 to -9.9	-14.6 to -11.1
Mean slope (SD)	0.7 (0.3)	0.7 (0.3)
Minimum and maximum location (dB SNR)	-14.6 and -7.6	-15.9 and -9.6
Difference from the grand mean (dB)	± 3.5	-3.0 – 3.4

3.4 Discussion

The current study attempted to equalise the words of the PAAST in quiet. The decision to conduct this study was made after looking at the results of the pilot, which showed high variability in locations for all the words in SIQ.

The locations of the words in the PAAST SiQ displayed a high range of variation (9.0 – 19.5 dBA) with an average location of 13.0 dBA. Some variation in the locations can be expected because, as mentioned in Section 2.2.3.2, the SiQ assesses audibility, and the SRTs (or location) are highly correlated with the PTA thresholds (Nilsson et al., 1994). Another reason could be that the chosen levels for the MoCS were insufficient to account for the variability of the intelligibility of words. This issue of possibly insufficient levels was addressed in the pilot study (6 levels), and additional levels were tested (a total of 9 levels) in Experiment 1. A, to ensure that the levels encompass the whole PF corpus. Even with the increased number of testing levels in the current study, some words such as HEN and WORMS were barely intelligible at the highest intensity level.

These findings suggested that the locations, at least for some words, were not accurate. Therefore, obtaining adjustment levels from these locations for equalisation would not be accurate, leading to false results when applying the test to patients.

Several studies (Summerfield *et al.*, 1994; Semeraro *et al.*, 2017; Al-Kahtani, 2020) showed that the MoCS was a suitable measure to assess the homogeneity of words in a speech test. The current study found that administering the PAAST SiQ using the MoCS led to poor intelligibility of some words even at the highest intensities. This effect could be overcome by increasing the levels of testing, requiring additional testing time, which could lead to exhaustion of the participants, possibly leading to inaccurate results. Another method to achieve equalisation of intelligibility of the words of the PAAST SiQ was chosen, namely the Interleaved Adaptive Procedure (ILAP). The ILAP allows for obtaining the parameters of each word without the restrictions of choosing testing levels, this is further discussed in Experiment 1.B in Chapter 4.

The results obtained from the PAAST SiN of the current study were similar to those obtained from the pre-equalisation stage of the PAAST SiN in the study by Al-Kahtani (2020). The similarity between studies suggested that the issue in equalisation stemmed from the type of testing background (noise versus quiet) and that the MoCS may have been suitable for assessing the homogeneity of the PAAST SiN but not SiQ because the latter resulted in high variability, even with increasing the number of levels for the MoCS.

3.5 Conclusion and Recommendations

The locations obtained from testing the participants with the PAAST SiQ using the MoCS were not reliable due to high variation in locations between words and the presence of erroneous locations of some words. A second experiment (Experiment 1.B) was conducted using the ILAP to obtain the parameters of the words of the PAAST SiQ and subsequently equalise the intelligibility of the speech material.

Chapter 4 Experiment 1.B Equalising the intelligibility of the speech stimuli of the Paediatric Arabic Auditory Speech Test (PAAST) using Interleaved Adaptive Procedure

4.1 Introduction

4.1.1 Background

Stimuli of a speech recognition test can vary in intelligibility leading to greater variability in measurements. To ensure reliable measures when using the speech test in an adaptive procedure, homogeneity of the test stimuli must be achieved (Leek, 2001).

The current experiment aimed to equalise the intelligibility of speech stimuli of the PAAST SiQ using an alternative method to the MoCS, which was considered a conventional method to achieve homogeneity of speech material (Summerfield *et al.*, 1994; H. Semeraro, 2015; Al-Kahtani, 2020). A different method was considered because, with the MoCS in Experiment 1.A, the range of differences from the mean location was wide (ranging from -4.05 to 6.51 dB). The importance of having “reasonable” ranges of values of differences from the mean location was because the levels of adjustment needed to equalise the intelligibility of the speech material of the PAAST SiQ are obtained from these values. Applying a wide range of adjustment levels could affect the words’ sound quality, rendering them too easy or hard to identify (Dietz *et al.*, 2014).

The interleaved adaptive procedure (ILAP) was considered for obtaining the parameters of the words of the PAAST SiQ in the current study. The rationale of choosing ILAP is explained in the next section.

4.1.2 Adaptive procedures

An adaptive procedure (AP) is a method used to measure a threshold, where the stimulus level on any trial is determined by the preceding stimulus and the participant’s responses (Levitt, 1971).

In an AP, the stimulus intensity level is decreased after a correct response and increased after a wrong one, thus creating reversals which constitute each of the two sides of the staircase, up and down (Leek, 2001). These reversals continue until they reach the limit set by the tester. The increment in which the stimulus is increased or decreased are referred to as steps, and a series of steps are considered a run. The level at which the test starts is called the initial level (Figure 4.1). The scored reversals, usually the ones obtained by the smallest step sizes, are averaged and that is considered the threshold (Zaltz et al., 2019).

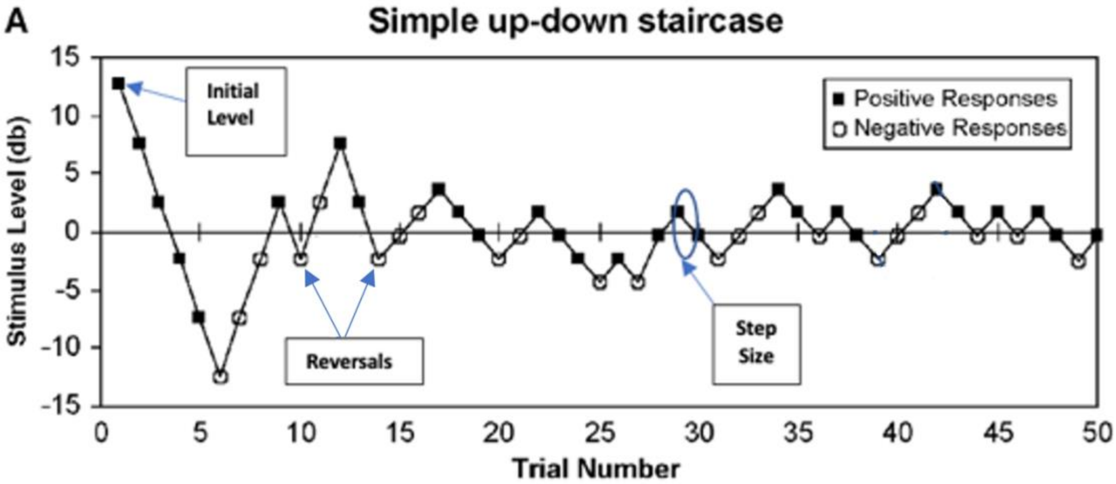


Figure 4.1 Adaptive track following a simple up–down 1 Down 1 Up staircase procedure (adapted from Leek, 2001)

Several parameters are set a priori to ensure effective threshold measurement including:

Down - Up rule

According to studies on AP, the number of correct responses required before decreasing the intensity of the stimulus would consequently affect the difficulty of the test, leading to an increased threshold at which the reversals are averaged (Levitt, 1971; García-Pérez, 1998). A general rule dictates that 1-down/1-up (1D/1U), 2D/1U, 3D/1U, and 4D/1U rules converge to SRT probabilities of 50%, 70.7%, 79.4%, and 84.1% respectively (Leek, 2001). Many studies, using the MTT (Summerfield *et al.*, 1994; Hall *et al.*, 2007) and the PAAST (Al-arfaj, 2018; Al-Kahtani, 2020), assessed the 70.7% SRT using the 2D/1U rule.

Step size rule

A general rule for step sizes, i.e., the increase and decrease of intensity for each reversal, is that the closer the trials are to the SRT, the more efficient the test is in identifying the threshold and calculating the slope from the tracks (King-Smith and Rose (1997) referenced by (Leek, 2001)). It is advisable to start with a large step size after the initial level (e.g., 8 dB) to home closer to the range of SRTs. The following proposed step size is half the first step size (4 dB), followed by a smaller step size (2 dB) that is continued until the set number of reversals is achieved (Leek, 2001).

Inappropriate step sizes may be problematic, with large step sizes possibly resulting in poorly placed data around the 50% point, and very small step sizes being time consuming leading to fatigue before reaching the SRT50. This can be avoided by setting large step sizes for the first two or three reversals and then running the remainder of the test using smaller step sizes.

Number of reversals

Reversals play an important role in measuring the threshold, because the latter is the average of the scored reversals, usually the last reversals with the smallest steps sizes. Therefore, more reversals lead to higher precision. The number of reversals in APs from studies varied between 40 and 4 reversals, with most studies using 6-10 reversals (García-Pérez, 1998). However, balance between the number of reversals and time should be considered, to avoid increasing the number of reversals to the point where fatigue or loss of attention can occur, especially in children.

Initial level

The initial level should be audible enough for the participant to give a correct response, but not too far from the threshold to allow the track to home in the range of the threshold after two or three trials, and preserve most of the testing time for determining the threshold within a small range of intensity levels (Levitt, 1971). The choice of the initial value is best taken based on prior knowledge from previous pilots or studies (Levitt, 1971). According to García-Pérez (1998), the effect of initial value on the convergence of the track to percent correct thresholds is mostly seen when the step size ratio is 1 (i.e. the up and down steps are equal).

Empirically, the AP has some potential advantages compared to MoCS owing to the former's ability to automatically concentrate the test trials within the dynamic range of the SRT (Levitt, 1971). The AP only requires well-chosen parameters, such as initial level and number of reversals to ensure accurate results (García-Pérez, 1998). Whereas in MoCS, all parameters of the psychometric function must be estimated concurrently, requiring a larger number of trials at well-distributed intensity levels. Testing the PAAST SiQ using the MoCS in Experiment 1.A showed that the intensity levels chosen were not sufficient to encompass the PF corpus of some words (e.g. HEN).

The MoCS has an advantage over AP in that, in the former, a PF is easily fit based on the thresholds obtained at each presented level, allowing for parameters such as the slope and location to be assessed. Yet, it is not impossible to obtain the slope from an AP track, though it would take more reversals (and time) to achieve (Leek *et al.,* 1991).

Most importantly, if the SRTs were the only measure needed to be assessed to achieve homogeneity, then AP (specifically interleaved adaptive procedure (ILAP)) might be a better choice than MoCS, especially with the PAAST SiQ.

The simple up-down AP with one track for all words would yield the SRT for the whole test for each individual (Leek, 2001), but not for each word. Therefore, an alternative method was proposed as a method to obtain the SRTs of each word, while also benefiting from the advantages of the AP. The method used in the current experiment is the interleaved adaptive procedure (ILAP).

Interleaved Adaptive Procedure

The method of ILAP was initially proposed as a way to follow changes in the PF during an experiment by tracing improvement in the performance of the participants, as well as signalling fatigue and distraction over the course of the experiment (Leek *et al.,* 1991). In the current experiment, the ILAP relies on using more than one adaptive procedure (tracks), one for each stimulus (word), resulting in an SRT for each word (Leek, 2001), in order to be able to assess and achieve homogeneity of the words.

4.2 Aims and objectives

The aim of this study was to equalise the intelligibility of the words of the PAAST SiQ using an ILAP.

The objectives directed towards achieving this aim were:

- Assess the effect of the words and repeats on the PAAST SiQ.
- Compare the SRTs obtained with the ILAP with those obtained from the MoCS.
- Identify the words that deviate from the mean SRT, and calculate the levels needed for adjusting the RMS amplitude of the words.
- Adjust the RMS amplitude of the words to achieve homogeneity between words
- Use Monte Carlo simulations (MCS) to assess the accuracy and precision of the PAAST SiQ when performed with certain ILAP parameters.

4.3 Method

4.3.1 Participants

The sample size was chosen according to previous studies that aimed to equalise speech stimuli (Summerfield *et al.*, 1994; Semeraro *et al.*, 2017; Al-Kahtani, 2020). The sample size decided on was n=20 Arabic-speaking NH adults, similar to that in Exp.1.A. Recruitment and inclusion criteria were the same as Section 3.2.2.2.

A total of n= 17 participants were tested in this experiment, the sample size was decreased from the set sample size n = 20 due to the COVID-19 lockdown. The average age of the participants was 31.25 (SD = 6.75) years, 9 were females and 8 were males, and all the participants' nationality was Saudi, except for two who were Yemeni.

4.3.2 Study design and procedure

Screening session

Same as Section 3.2.2.3.

Testing session

The PAAST SiQ was applied through MATLAB software using an in-house MATLAB code specific for this study. The test consisted of a graphical interface containing 14 words of acoustically similar Arabic names, and headphones (Sennheiser HD 650) connected to the computer. The participants faced a monitor, with the interface containing the test words on display (Figure 4.2).

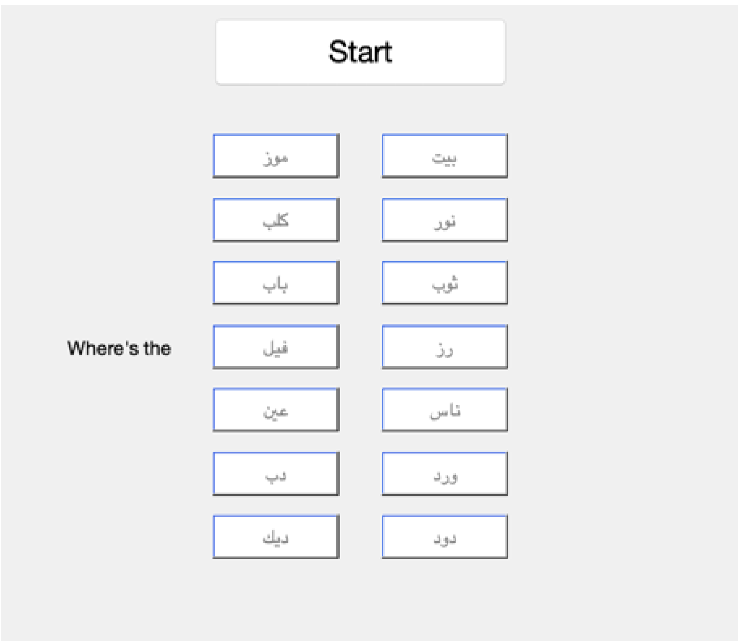


Figure 4.2 Graphical interface of the PAAST SiQ - Pre-equalisation stage

Participants were asked to read the words displayed to ensure they understood them. They were instructed to listen to the sentences (‘where is the’ + name) binaurally through the headphones connected to the computer and click with a mouse on the corresponding name on the screen. They were advised to guess if they did not hear the word or if they were not sure what the word was, they were also encouraged to choose the closest word to what they may have heard. The participant’s responses were automatically recorded on the computer.

The ILAP was applied in the form of 14 tracks that were interleaved in an entirely random fashion. Each track was designated for a specific word, so that the participant could be listening to a word (1: BANANA) from track (1), and the second stimulus would be word (4: ELEPHANT) from track (4) and so forth.

The ILAP was applied in a 1-down 1-up procedure, resulting in an SRT that would correspond to approximately 50% (SRT50) (Shen, 2013). The initial level (IL) was 30 dB A, and the initial step size was 8 dBA. The second step was 4 dB and afterwards all following steps were a 2 dB step size, up until a total of 8 reversals were completed (a maximum of 70 trials for each track). The SRTs were calculated from the mean scored reversals of the last 6 reversals, meaning that the stimulus levels corresponding to each of the reversals were averaged resulting in an SRT. The parameters are summarised in Table 4.1. The resulting SRTs for each participant were a total of 14 SRTs, one for each word (track).

Table 4.1 Parameters used for the ILAP in the pre-equalisation stage

Parameter	**Condition N**
Up-down rule	1 down 1 up
Step size rule (dB)	8,4,2
Reversals for each step	1,1,6
Initial value	30 dBA

The PAAST SiQ was applied twice, and results of both repeats were averaged to obtain the levels of adjustments. Each repeat took approximately 20 minutes. The total testing time was approximately 45 minutes including breaks.

4.3.3 Materials, calibration, and safety

Same as section 3.2.2.6. This experiment was conducted in the period leading up to the declaration of COVID-19 as a pandemic, and the data collection was terminated once lockdown was enforced.

4.3.4 Ethical considerations

Ethical approval (ERGOII 46958.A1) for this experiment was granted from the University of Southampton on 16 April 2019 (Appendix I.3). Ethical approval for the application (569-18) to King Abdulaziz University was approved on 13 November 2018 (Appendix I.2).

4.3.5 Analysis Strategy

The analysis included descriptive and statistical analysis directed towards obtaining the results of each objective.

Assessing the effect of the words and repeats on the PAAST SiQ

Descriptive analysis displayed the mean and standard deviations (SD) for each word across participant in repeat 1, 2 and the average of both repeats (Repeat Avg.).

The effects of the words and the repeats on the SRTs were analysed by applying 2-way repeated measure ANOVA (RM-ANOVA). The two factors were words (14 levels) and repeats (2 levels), and the dependent measure was SRT.

Comparing the SRTs obtained with the ILAP with those obtained from the MoCS.

Descriptive analyses were used to compare between the results of the PAAST SiQ obtained from MoCS from Experiment 1.A (Section 3.3) with those obtained with ILAP.

Adjusting the RMS of the words to achieve homogeneity

The differences of the grand mean SRT (averaged across words, participants, and repeats) from the mean SRT of each word averaged across repeats (Repeat Avg.) were calculated to obtain the levels of adjustment. These levels are applied to the RMS of the amplitude of each word to equalise the intelligibility of the speech material. It was suggested that a difference range larger than 2.5 dB would indicate adjustment of the words. This range was chosen based on the fact that it was expected to have a larger range of SRTs in SiQ compared to those in SiN (Summerfield *et al.*, 1994).

Monte Carlo simulations (MCS)

A MCS was performed to assess the precision and accuracy of the parameters of the ILAP to ensure the use of the most effective parameter values and options for obtaining SRTs in future experiment. The parameters and characteristics tested are listed in Table 4.2.

Table 4.2 Parameters and characteristics tested in the simulation

Parameters					**Characteristics**	
initial level (IL) (dBA)	**Maximum level (dBA)**	**Reversals (R)**	**Step size (dB)**	**Down-up rule**	**Location (L) (dBA)**	**Slope (S)**
20	50	1,1,6	8, 4, 2	1D1U	6	0.1
40		1,1,12			12	0.5
					18	0.9

Parameters highlighted in grey were fixed

The main observations of interest in the simulation were the accuracy and precision. Accuracy of a measurement procedure refers to how close the values of the mean SRTs are to the true expected value (Zaltz *et al.*, 2019), this was assessed by calculating the difference between the mean SRTs and the true threshold ($V^{True} – V^{Estimate} = V^{Diff}$). Smaller V^{Diff} indicate higher accuracy. Precision is how similar the mean SRT values were across all conditions, in other words, the variability of the estimate, which was assessed by the SD. The smaller the SD value, the more precise the results are. Details of the MCS simulation methods and analysis are mentioned in Appendix F.2.

All statistical analyses were done using SPSS (v.27).

4.4 Results

4.4.1 Descriptive results

The mean SRT was lower in Repeat 2 (mean = 10.17 dBA, SD = 3.16 dB) compared to repeat 1 (mean = 11.31 dBA, SD = 3.01 dB). The words with the highest SRTs were HEN (mean = 16.91 dBA, SD = 5.32 dB) followed by WORMS (mean = 16.35 dBA, SD = 3.06 dB). The words with the lowest SRTs were EYE (mean = 7.10 dBA, SD = 3.33 dB) followed by DOOR (mean = 7.32 dBA, SD = 3.89 dB). The mean SRT of each word and each repeat, as well as the average SRTs across repeats are displayed in Table 4.3. A Boxplot of the average SRT of Repeat 1 and 2 for each word is displayed in Figure 4.3.

Table 4.3 Mean SRTs and standard deviation (SD) of each word in repeats 1, 2 and average 1&2

Words	Repeat 1		Repeat 2		Average Repeat 1&2	
	Mean (dBA)	SD (dB)	Mean (dBA)	SD (dB)	Mean (dBA)	SD (dB)
BANANA	12.45	2.49	11.64	2.64	12.04	2.41
DOG	8.76	3.24	8.30	2.71	8.53	2.52
DOOR	7.88	4.04	6.76	3.89	7.32	3.89
ELEPHANT	11.45	2.14	11.49	2.73	11.47	2.32
EYE	7.35	3.37	6.86	3.69	7.10	3.33
BEAR	8.81	3.78	8.32	2.92	8.57	3.29
HEN	16.94	5.02	16.89	5.90	16.91	5.32
HOUSE	9.90	2.79	9.72	2.83	9.81	2.76
LIGHT	13.90	3.80	12.28	2.89	13.09	2.85
MENDRESS	11.06	2.06	10.01	2.74	10.53	2.29
RICE	13.84	3.13	12.39	2.70	13.12	2.74
PEOPLE	10.69	4.12	9.63	4.49	10.16	4.25
FLOWER	9.05	3.42	7.93	2.51	8.49	2.77
WORMS	16.24	3.19	16.45	4.09	16.35	3.06
Mean	11.31	3.01	10.17	3.16	10.964	3.08

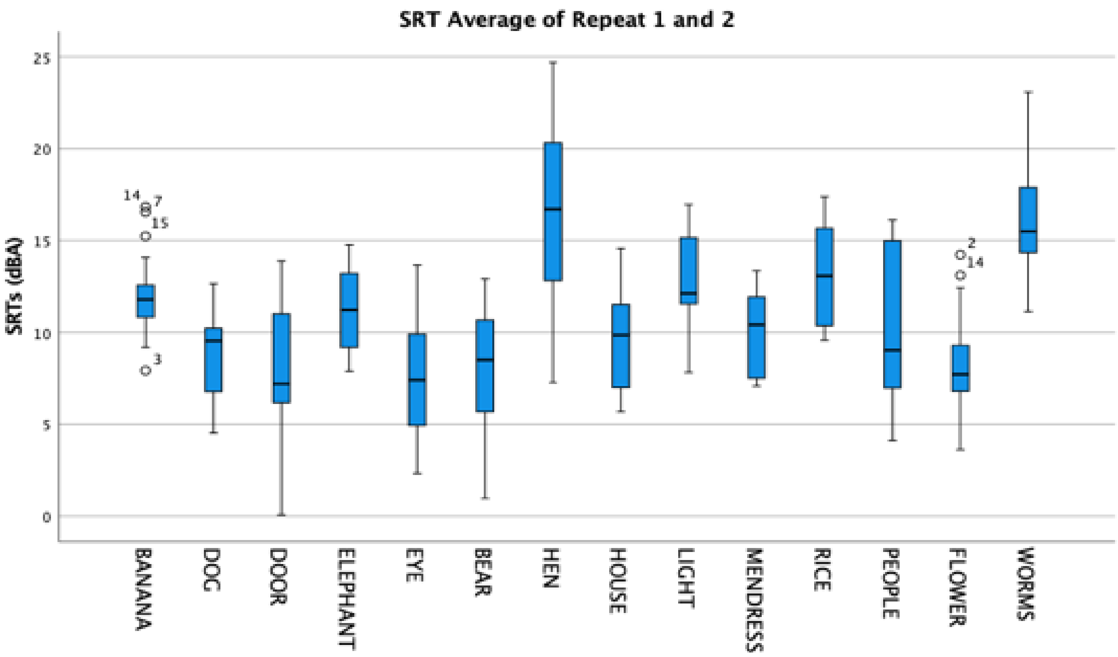


Figure 4.3 Boxplot of the average SRTs across repeat 1 and 2 for each word for all participants (n = 17)

4.4.2 Effect of words and repeats on the words of the PAAST SiQ

Details of the 2-way RM-ANOVA are included in Appendix F.1. The main results were:

1- **Effect of words:** There was a statistically significant difference between the SRTs of the words ($p < 0.001$). The words that had significantly different SRTs than most other words were WORMS followed by HEN.
2- **Effect of repeats:** Mean SRTs across all words exhibited a significant decrease of 0.69 dB (SE = 0.17, 95%CI = -1.05 to -0.33) in repeat 2.

4.4.3 Comparison between the results of the PAAST SiQ obtained using MoCS and ILAP

The SRTs obtained from the MoCS (Experiment 1.A) were compared with those obtained from the ILAP (current study). Generally, the mean SRTs of each word from the current study were lower than those obtained using the MoCS, which could be either due to the difference in procedure or difference between participants from each experiment. There was a general decrease in the SDs of the SRTs of each word in the current study compared to those obtained by MoCS. The biggest difference in SDs was seen with the word WORM (SRT (MoCS) = 7.77 dB, SRT (ILAP) = 3.06 dB, diff. = 4.71 dB). There was an improvement in the grand mean SRTs with ILAP compared to MoCS (mean = 10.96, SD = 3.08 and mean = 12.62, SD = 4.57 dB, respectively), but the differences between the mean SRTs of each word and the grand mean were similar in both methods (Figure 4.4). Table 4.4 displays this comparison in detail.

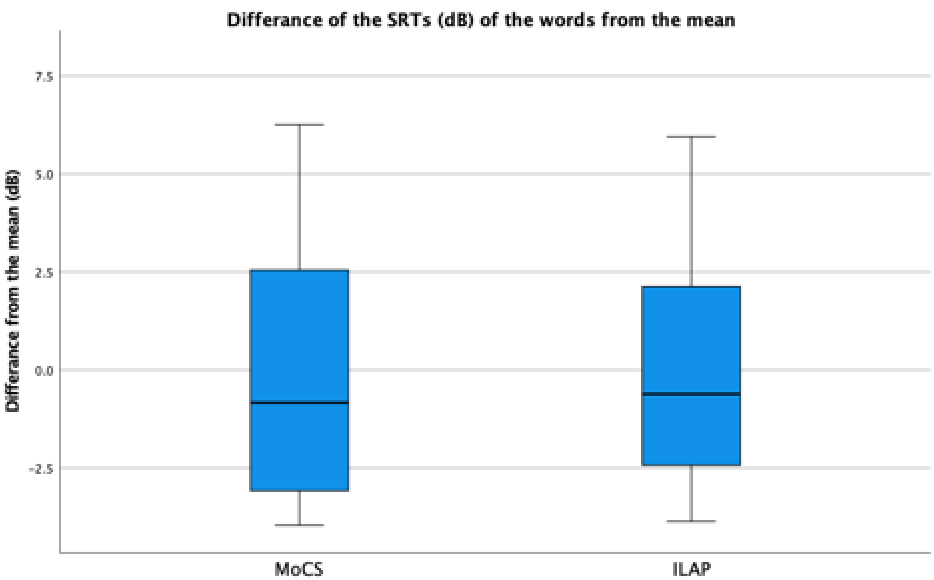


Figure 4.4 Boxplot of the differences of the grand mean SRT from the mean SRTs of the words (n = 14) using the MoCS and ILAP.

Table 4.4 Comparison between mean SRTs and Standard deviations obtained using the MoCS (Experiment 1.A) with those obtained from the current Experiment using ILAP

Words	Method of Constant Level Stimuli (n=20)			Method of Interleaved AP (n=17)			Difference between SD of MoCS and ILAP
	M (dBA)	SD (dB)	Difference of SRTs from mean	M (dBA)	SD (dB)	Difference of SRTs from mean	
BANANA	14.76	5.35	2.14	12.04	2.41	1.08	2.94
DOG	11.20	2.61	-1.41	8.53	2.52	-2.43	0.09
DOOR	8.68	3.82	-3.94	7.32	3.89	-3.64	-0.07
ELEPHANT	12.64	3.43	0.03	11.47	2.32	0.51	1.11
EYE	8.65	3.10	-3.96	7.1	3.33	-3.86	-0.23
BEAR	9.53	3.39	-3.08	8.57	3.29	-2.39	0.10
HEN	18.87	6.39	6.26	16.91	5.32	5.95	1.07
HOUSE	10.94	3.35	-1.67	9.81	2.76	-1.15	0.59
LIGHT	15.16	4.67	2.55	13.09	2.85	2.13	1.82
MENDRESS	12.35	5.77	-0.26	10.53	2.29	-0.43	3.48
RICE	16.20	5.64	3.59	13.12	2.74	2.16	2.90
PEOPLE	10.88	4.35	-1.74	10.16	4.25	-0.80	0.10
FLOWER	9.12	4.38	-3.49	8.49	2.77	-2.47	1.61
WORMS	17.62	7.77	5.00	16.35	3.06	5.39	4.71
Grand Mean	12.62	4.57		10.96	3.08		1.49

4.4.4 Adjustment levels for equalising the RMS of the words

The differences between the mean SRT for each word and the grand mean were calculated and displayed in Table 4.5. The differences were rounded to the nearest one in order to obtain the levels required for adjusting the RMSs of the words. The range of adjustment levels was -4 to 6 dB.

Table 4.5 Adjustment levels needed for each word rounded to the nearest 1 dB

Word	Adjustment level (Rounded to the nearest 1 dB)
BANANA	1
DOG	-2
DOOR	-4
ELEPHANT	1
EYE	-4
BEAR	-2
HEN	6
HOUSE	-1
LIGHT	2
MENDRESS	0
RICE	2
PEOPLE	-1
FLOWER	-2
WORMS	5

The words with an adjustment level > 4 are highlighted in grey

4.4.5 Monte Carlo simulations

Details of the Monte Carlo simulations are included in Appendix F.2. The main results showed that:

1- **The slope:** the slope is one of the characteristics of the PF of a test. The steeper the slope was (0.9 or 0.5), the more similar the results were to the true value (V^{Diff} <0.2 dB), as opposed to the shallow slope of 0.1 where V^{Diff} differed up to 2.5 dB, suggesting that steep slopes lead to more accurate results. Precision was best at slope 0.9 and 12 reversals (SD = 0.7 dB).

2- **The Initial level:** The initial level per-se did not affect the results of the adaptive procedure, but rather the distance between the initial level and the location (IL-SRT), and that effect was only evident when the slopes were shallow. The smaller the IL-SRT, the more accurate the results were (down to $V^{diff.}$ = 0.5 dB).

3- **Reversals:** Accuracy (V^{Diff}) and precision (SD) were improved by 1 dB and 0.4 dB, respectively, with larger number of reversals (12 reversals) compared to 6 reversals, even in conditions where the slopes were steep (0.9).

4.5 Discussion

Effect of words and repeat on the PAAST SiQ using ILAP

The results of the RM-ANOVA showed significant differences between the SRTs of the words. The pairwise comparisons signalled out two words with SRTs significantly different from most words, namely WORMS and HEN, indicating these words could potentially be problematic and might have a higher value of difference from the mean compared to the other words. This result impacted decisions regarding adjusting the RMS amplitudes to achieve homogeneity of the speech stimuli, which is discussed later in this section.

The analysis showed that the mean SRTs differed between repeats, with a significant decrease of 0.69 dB meaning that the participants found the words easier to hear in Repeat 2. This analysis examined a single aspect of test-retest reliability which is systematic error. Systematic errors are errors resulting from a specified factors such as a learning effect (Lovett *et* al.,2013). Previous studies examined the reliability of the MTT (Lovett *et* al., 2013) and the PAAST (Al-arfaj, 2018; Al-Kahtani, 2020). Similar improvement in performance was observed with the study by Al-arfaj (2018) which assessed test-retest reliability of the PAAST SiQ.

Comparison between ILAP and MoCS and adjustment levels

Equalising the intelligibility of speech material is considered a requirement before applying a speech test to ensure accurate results. The equalisation of speech material in the current PhD was performed over several trials, primarily due to issues observed with the MoCS in Experiment 1.A including the poor intelligibility of words at the highest testing levels, warranting the decision to use the ILAP to obtain measurements for adjustment.

Summerfield *et al.*, 1994 obtained the levels needed for adjustment using the MoCS, which were -3.5 to 5.9 dB in quiet and -2.5 to 2.5 dB in noise, indicating large threshold ranges in quiet compared to noise. These findings were similar to the case of the PAAST when applied in MoCS in Experiment 1.A, where the range of locations in noise (−15.9 to −9.6 dB SNR) was smaller than that in quiet (8.91 to 19.47 dBA). In the MoCS, the higher end of the range of the locations of the words (19.5 dBA) was close to the highest level presented (20 dBA). Whereas in ILAP, the highest SRT for a word was 16.9 dBA.

Another advantage of using the ILAP was the time factor. Testing using the ILAP took approximately 45 minutes, whereas with the MoCS, it took 75 minutes to complete the test. the increased time could lead to fatigue which might affect accuracy.

These observations suggest that even though the adjustment levels when obtained by ILAP and MoCS were similar, ILAP can be performed in a shorter time than with MoCS and does not require deciding levels of testing a priori.

Adjustment of the RMS of the PAAST SiQ

The difference from the mean for all words ranged between -3.86 and +2.15 dB, except for two words: HEN and WORMS, which had a difference of more than +5 dB. Adjusting the RMS by increasing it by ≥ 5 dB could affect the natural sounding of the words, consequently affecting the performance on the test. Several studies set a limit to the adjustment levels to preserve the natural sounding of the speech material. For example, Dietz *et al*. (2014) set the maximum adjustment level to ±3 dB. In their study, Hochmuth *et al*. (2012) set the level to ±8 dB as the maximum acceptable difference from the mean.

A similar approach was taken with the equalisation of the words of the PAAST SiQ, where the adjustment level was set to ±4 dB, and pairs containing words with SRTs outside that range (namely, HEN and WORMS) were eliminated. The eliminated pairs were (ELEPHANT- HEN and LIGHT – WORMS). The RMS levels of the words in the remaining five pairs were increased or decreased based on the adjustment level of each word using MATLAB. The post-equalisation stage, where homogeneity of the words was assessed after the RMSs of the 10 words has been adjusted, was performed as part of Experiment 2 (Section 5.5.1).

It can be argued that elimination of words may affect the efficacy of the test, but as mentioned in the literature review, the MTT, which the PAAST was adapted from, can be applied using down to three pairs and still can provide information about hearing loss in children with OME (Hall *et al.,* 2007). In their study to assess the test-retest reliability of the AMTT in noise in hearing impaired children, Lovett *et al.,* (2013) eliminated 2 pairs of words because they were difficult for children to identify.

Rationale behind eliminating pairs and not single words in the PAAST SiQ

As mentioned in Chapter 2, the PAAST SiQ was adapted from the AMTT. The MTT/AMTT was originally designed as a discrimination test to ensure children could recognise the heard words and choose their corresponding objects, even in the presence of another object with an acoustically similar name. Therefore, studies using the MTT and AMTT followed the same procedure in dealing with children who were unfamiliar with certain words, by eliminating the hard word and its pair, (Hall *et al.,* 2007; Lovett *et al.,* 2013) to avoid having a single word acoustically different from the rest, which could affect test accuracy.

It is unknown whether there was a pattern in the AMTT in which a listener would incorrectly choose a word when the corresponding word in the same pair is presented, for example, incorrectly choosing “CUP” when the presented word was “DUCK”. The current experiment assessed the differences between the SRTs of the words of the PAAST SiQ but did not provide information about specific patterns of incorrectly choosing one word of the pair instead of the other. This would require analysing the patterns of incorrect responses to investigate if the listener had difficulty discriminating between the two words of one pair at very low intensities.

However, looking at the word pairs displayed in Table 4.6, differences are found in beginning and ending consonants for all word pairs, while other words not within the same pair share similar characteristics, such as the word MENDRESS (/ θɔb/) and BEAR (/ dub/), where the /ɔ/ in MENDRESS is a single long sound, whereas /u/ is a short vowel, but they both share the same consonant at the end of both words (/b/). A similar pattern is seen with the words DOOR (/ bab/) and DOG (/ kʌlb/). It is therefore possible that a listener would incorrectly choose DOOR when the word presented is DOG, even though they are not within the same pair.

Table 4.6 The words of the PAAST and their phonetic transcription

Pair	**Word**	**Phonetic Transcription**	**Word**	**Phonetic Transcription**
1	Door	/ bab/	People	/ nas/
2	Elephant	/ fil/	Hen	/ dik/
3	House	/ bet/	Eye	/ ʕen/
4	Mandress	/ θɔb/	Banana	/ mɔz/
5	Light	/ nuɽ/	Worms	/ dud/
6	Bear	/ dub/	Rice	/ ɽuz/
7	Flower	/ wʌɽd/	Dog	/ kʌlb/

Highlighted in grey are the words that have been eliminated post-equalisation

Therefore, it is advisable to treat this test as a speech recognition test (as done with the AMTT (Hall *et al.,* 2007)), because the listener should be able to recognise the words, but at the same time it must be ensured that the elimination of any word is accompanied with elimination of its pair, to avoid having a single word stand out, rendering it easily recognisable, and possibly affecting the accuracy of the test.

Monte Carlo simulations (MCS)

The MCS suggested that the accuracy of SRTs was at its best when the slopes were steep or medium, regardless of the initial levels, reversals, or location. In the shallow slopes, where the accuracy started to be affected, the SRTs were closest to the true threshold in the condition with the largest location (L = 18 dB), the smallest IL (20 dB) and the largest number of reversals (R = 12). The initial level per-se did not affect the results of the adaptive procedure, but rather the distance between that level and the threshold (IL-SRT), and that effect was only evident when the slopes were shallow. Similar findings were observed when assessing precision with SDs, but the precision of obtaining the SRT can be affected by the number of reversals even in the steep slopes, where the SD improves by up to 0.4 dB with 12 reversals compared to six.

Zaltz *et al.* (2019) aimed to assess the effect of a choosing a logarithmic versus linear change in step size in an adaptive procedure of a frequency discrimination task. Similar to the current study, they varied some parameters (level of performance, level of attention lapses, and adaptive rule), but fixed the slope to 0.5 based on previous human experiments. They found similar results with both adaptive rules, but a tendency for a better accuracy but less precision with the linear rule. They concluded that the small step size played a role in the similarities in accuracy between all conditions, and that the larger step sizes obtained with the logarithmic rule might have affected the accuracy. Another possible reason for the similarities between all conditions was the choice of the medium slope which played an important role in precision and accuracy as was evident in the current study.

In their study, Watkins *et al*. (2016) measured the effect of presentation level (initial level) on the efficiency of a speech in noise test. The simulations were run in three different procedures including an ILAP where the test was presented at three different levels of presentations (IL). They found no difference in SDs in presentation levels of noise levels 65- and 80-dB SPL (with varying noise levels based on participants' responses) on the SRTs. Only when the presentation level was 50 dB SPL did the SRTs differ significantly due to inaudibility to the IL. These results further support the idea that initial levels that are not too far from the threshold are sufficient to obtain accurate and precise thresholds, and that a very low initial level (lower than the highest possible thresholds) can result in straying away from the range of threshold, and the last small step sizes would be wasted on levels far from the threshold, rendering the results inaccurate.

The simulations allowed controlling for location and slope, and when the conditions included the mean location (12 dB) and slope (0.4) from Experiment 1.A, the yielded SRTs were high in accuracy and precision regardless of all other parameters.

Choice of parameters for Experiments 2 and 4

In Experiment 1.B, the initial level was 30 dBA, which was considered a reasonable level because it was far enough form the highest SRT (17 dBA) and was considered audible. This level was chosen for the post-equalisation stage in Experiment 2 to ensure homogeneity. Considering Experiment 2 involved lengthy testing duration in adults, where participants were asked to perform the PAAST SiQ in an ILAP in three conditions (normal and two SCHL), it was decided to keep the 6 reversals parameter.

In Experiment 4 where children with OME were tested, the initial level was chosen at 50 dBA, to include all possible SRT71 (2D1U), which were expected to be high in OME ears (and in young children). The number of reversals chosen was 6 because the time and attention factors could be accentuated in children. Previous studies using the AMTT on young children support the use of 6 reversal to obtain SRT71 (Summerfield *et al.*, 1994; Hall *et al*., 2007; Lovett *et al*., 2013).

4.6 Conclusion

This chapter was directed towards equalising the words of the PAAST SiQ and assessing the effects of parameters on the thresholds using the ILAP. The main conclusions were:

- Applying the PAAST SiQ in an ILAP method resulted in less variable SRTs, and was performed in a shorter time, compared to testing with MoCS.
- Problematic words were identified, and the decision was made to eliminate the two pairs containing them (ELEPHANT- HEN and LIGHT – WORMS), to preserve the natural sounding of the words. The remaining 10 words' RMSs were adjusted for intelligibility.
- Computer simulations (MCS) showed that the number of reversals play a role in precision even in words with steep PF slopes. The IL should be audible, but at the same time not far from the estimated threshold to ensure accuracy of the results.

Chapter 5 Experiment 2: Measuring the effect of OME-related simulated conductive hearing loss on the PAAST SiQ

5.1 Introduction

The main aim of this study was to test the sensitivity of the PAAST SiQ to OME related simulated CHL (SCHL). A pilot of this study was conducted in the beginning of the PhD (Chapter 3), the results of the pilot indicated that the words of the PAAST SIQ were not equally intelligible, even though the same words were equalised for intelligibility in noise (Al-Kahtani, 2020). The pilot study was revisited again in Experiment 2 after the words of the PAAST SiQ were equalised for intelligibility, to achieve Aim 2 of the PhD.

The sensitivity of each word to the different intensities of speech stimulus has not been investigated in either the English version of the MTT or the Arabic version (PAAST). McCormick (1977) suggested that if a child was not familiar with all toys (words), the test can still be carried out with fewer pairs. Later studies used as few as 3 pairs of toys (Hall *et al*., 2007; Lovett *et al*., 2013). The aim of this study was to measure the sensitivity of the PAAST SiQ to SCHL in adults, and to investigate whether all words were affected equally by OME-related SCHL.

The current study also assessed the adjusted speech stimuli from Chapter 4 to verify that all words had similar SRTs. The post-equalisation step was necessary to ensure that any variations in SRTs using the PAAST SiQ in future studies would be due to factors related to the participants or hearing conditions rather than factors related to the speech material.

5.1.1 Hearing loss simulation

5.1.1.1 Conductive hearing loss simulation

Hearing loss simulation (HLS), if done accurately, is an effective way to study HL in isolation (Moore and Glasberg, 1993). This would be especially helpful in measuring the effect of OME on SRTs using a test such as PAAST SiQ because HL associated with OME has specific audiogram configurations that usually affects certain frequencies (Cai and McPherson, 2016). Like any simulation study, there are limitations with HLS. As accurate as HLS may be designed to mimic a certain hearing loss condition or type, it may not always convey the true hearing loss experienced by patients, and it may not take into account the variations that can be found between patients with actual HL.

There are several types of hearing loss simulation methods including threshold elevation, loudness recruitment and reduced frequency selectivity (Blyth, 2019). One or more HLS methods can be used depending on what needs to be studied as well as the type of test used.

Several studies investigated the effect of OME-related SCHL in individuals with NH. These studies used the threshold elevation method to simulate this type of HL. The reason why threshold elevation was the best method to simulate this type of HL is because CHL caused by OME is characterised by decreased audibility caused by elevated hearing thresholds (Plomp, 1978).

The simplest way of achieving threshold elevation is through inserting an ear plug into the ear canal (Adelman *et al*., 2015). In their study, Adelman *et al*., (2015) assessed AC, BC, and soft tissue conduction (STC) using a bone vibrator in simulated CHL and SNHL. The SCHL, which was achieved by an ear plug, elicited AC threshold elevation of 21–37 dB across listeners, but no change was seen with BC or STC. The AC threshold elevation was consistent with the expected OME-related HL, which was slight to mild, but the AC elevation was more evident in high frequencies (Adelman *et al*., 2015), which may less commonly occur in some cases in CHL associated with OME (Cai *et al.*, 2018). Another recent study by Snapp *et al*. (2020) used an ear plug in the external ear canal to study the effect of SCHL on localisation, and the benefit of using Bone Conduction systems (BDS) on localisation in SCHL.

Other studies used methods that adjusted the stimulus intensity levels. A study by Cai *et al*. (2017) matched a group of children who were otologically normal (no HL), with children with CHL caused by OME, the pair were sex- and age- matched, and the difference in their PTA thresholds at frequencies from 125 to 8000 Hz was designated as the attenuation level in that pair. Threshold elevation in this study was achieved by reducing the intensity level of the speech material by filtering the Mandarin Hearing in Noise Test (HINT) through a graphic equaliser. A study by Penn *et al*. (2004) used a similar method of simulation, digitally filtering words from the Phonetically Balanced Kindergarten (PBK) word lists and attenuating the levels of intensity in these filters, and applying the tests on children and adults in normal (no attenuation) and attenuated conditions. Their study allowed for simulating the commonly seen audiograms in children with OME related CHL, where thresholds were elevated at low frequencies and gradually improving until reaching the 2kHz frequency where thresholds were the lowest (better). Findings in the study by Penn *et al*. (2004) showed a significant elevation of thresholds of the low frequencies in simulated conditions, and significantly better performance in adults compared to children.

The HLS chosen for the current study was threshold elevation, which was designed by a supervisor (Dr. Daniel Rowan) at the University of Southampton. The HLS software has been tested and was based on the method provided by Moore and Glasberg (1993). The same software was used by a former SOTON PhD student (Semeraro, 2015), to simulate hearing loss using both threshold elevation and loudness recruitment to measure the effect of simulated SNHL on speech intelligibility.

Hearing loss simulation using threshold elevation was achieved by processing the audio files of the words through a simulation code specifically designed for MATLAB. The version used to run the HLS code for this experiment was MATLAB (R2020a). The process of threshold elevation using this software was achieved by implementing filters digitally, where each filter had a central frequency. These signal were then attenuated, time aligned and mixed to produce an output representing the simulated stimulus (Moore and Glasberg, 1993). This process resulted in amplitude reduction in certain frequencies (low frequencies) of each word.

One might argue that decreasing the stimulus level would automatically lead to the inability to hear the words, leading to threshold elevation. The current study aimed to assess whether the audiogram's configuration would lead to increased SRTs and whether preserving high frequencies would lead to some words with high-frequency information being unaffected by the simulation.

5.1.1.2 Applying the Conductive hearing loss simulation

Three SCHL audiograms with rising hearing loss configuration were chosen based on common, as well as less common configurations (Cai and McPherson, 2017). These conditions differed in the frequency knee point at which the thresholds at the designated frequency and the proceeding frequencies were left unattenuated, while the thresholds of the frequencies preceding the knee point frequencies were attenuated. Condition N was the normal, unattenuated audiogram, whereas Conditions 2K, 4K and 8K simulated rising CHL with a knee point of 2 kHz 4 kHz and 8 kHz, respectively. An illustration of the proposed audiogram is shown Figure 5.1. Table 5.1 demonstrates the intensity levels (dB HL) at each frequency in each condition.

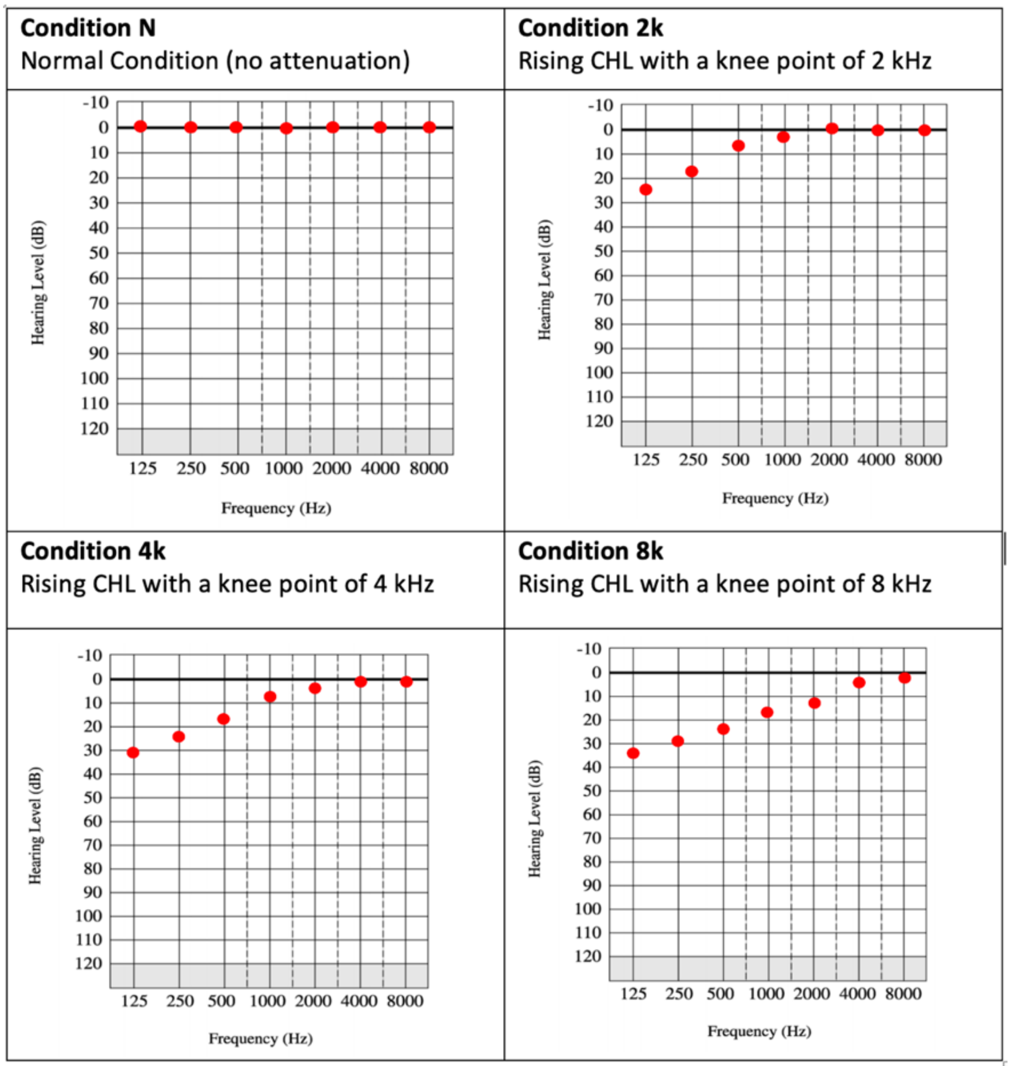


Figure 5.1 Proposed audiometric configurations of the threshold shift for the SCHL in the current study

Table 5.1 Thresholds shifts (elevation) in the conditions.

	0.125 kHz	**0.25 kHz**	**0.5 kHz**	**1 kHz**	**2 kHz**	**4 kHz**	**8 kHz**
Condition N	0	0	0	0	0	0	0
Condition 2k	24	18	9	3	0	0	0
Condition 4k	30	24	18	9	3	0	0
Condition 8k	33	30	24	18	9	3	0

Condition "N" with no simulation, "Condition 2K" with threshold elevation at frequencies preceding 2kHz, "Condition 4k" with threshold elevation preceding 4KHz, and "Condition 8k" with threshold elevation preceding 8kHz

The degree of the rise was not consistent and varied between 3, 6 and 9 dB/octave, with the worst threshold not exceeding 35 dB HL. This maximum level (35 dB) was acceptable because according to a systematic review on hearing loss in children with OME, the average PTA thresholds (500, 1000, and 2000 Hz) can reach up to 35 dBHL (Cai and McPherson, 2016). The degree of the slope (i.e. change in thresholds across frequencies) did not pose a problem because this experiment aimed to test the effect of the rising audiometric configuration on the PAAST and compare the results with a non-attenuated (flat) audiogram.

The same audiograms in the current study were used in the Pilot study where the mean locations of the PF (dBA) across the words of the PAAST SiQ in all subjects were plotted in boxplots (Section 3.2). Observations from the results of the experiments showed a general increase in mean locations in condition 2K compared to condition N (no simulation, flat audiogram), and an increase in mean location in condition 4K compared to 2K (Figure 5.2).

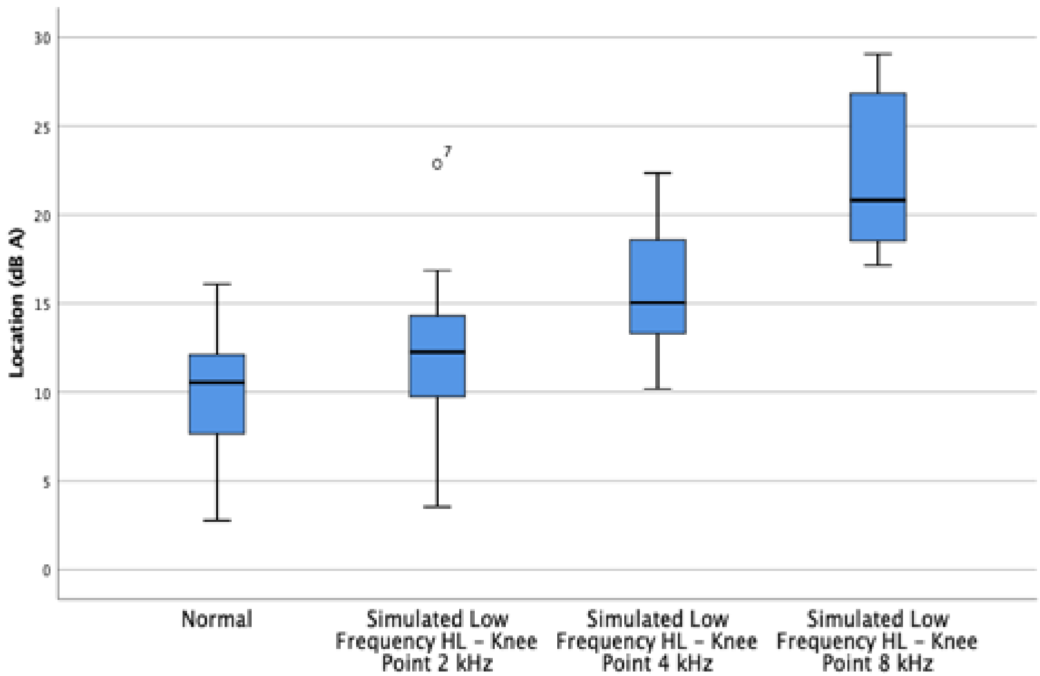


Figure 5.2 Boxplots of the average location across the words in each subject (n = 20) (i.e. 20 points in each boxplot) in the four hearing conditions (Pilot study – Chapter 3)

Based on these results, conditions 2K and 4K as well as condition N were chosen for the current experiment. The decision to eliminate condition 8K was based on two reasons. Firstly, because condition 2K is more commonly seen in children with CHL caused by OME, followed by condition 4K (Cai and McPherson, 2017). Secondly, testing all four conditions would prolong the test, which could lead to fatigue, possibly affecting performance and accuracy of the results.

5.2 Research question, aims and objectives

5.2.1 Research Question

Is the PAAST SiQ sensitive to SCHL associated with OME

5.2.2 Aims

1. To ensure all the words of the PAAST SiQ were equal in intelligibility (post-equalisation)
2. To ensure that the words of the PAAST SiQ are sensitive to SCHL associated with OME
3. To investigate whether the words of the PAAST SIQ are equally sensitive to SCHL associated with OME

5.2.3 Objectives

The aims of this study were achieved through the following objectives:

1. Recruiting a sample of Arabic-speaking otologically normal adult participants.
2. Using SCHL to recreate the low frequency rising audiogram that typically occurs with OME. These conditions were Condition N, 2K, and 4K.
3. Obtaining the SRT of each of the 10 words for each participant in all conditions using the PAAST SiQ in an ILAP.
4. Assessing the homogeneity of the words of the PAAST SiQ post-equalisation from the SRTs obtained from Condition N.
5. Comparing the mean SRTs and change in SRTs of all 10 words across the three conditions.

5.3 Method

5.3.1 Study design

Data collection was carried out in the Audiology clinics at King Abdul-Aziz University Hospital (KAAUH) in Jeddah between December 2020 and February 2021. Participants underwent testing of their hearing abilities using the PAAST SiQ. All testing was performed in a soundproof room, with a background noise level no greater than 30 dBA.

The experiment lasted approximately 80 minutes and was composed of three sessions. Session 1, a screening session, which lasted 10 minutes, and sessions 2 and 3, the main testing sessions, which lasted 35 minutes each. All participants attended sessions 2 and 3 on the same day and were given the opportunity to take a five-minute break every 20 minutes during testing.

In the current study, the method used to obtain the SRTs was the ILAP, which was the same method used in Experiment 1.B (Section 4.3.2), to obtain the SRTs of each word in the PAAST SiQ. The method of ILAP, with the right set of parameters, can provide precise SRTs (Levitt, 1971) in less time compared to applying the test using the MoCS. The procedure in the current study lasted 80 minutes, whereas the same procedure, which was done in the Pilot study (Section 3.2), using the MoCS lasted approximately 150 minutes.

The ILAP test was applied in a 1-down 1-up procedure, which would result in obtaining an SRT that would correspond to approximately 50% (SRT50) (Shen, 2013).

The parameter values were chosen based on the results of the MCS which was part of Experiment 1.B (Section 0). The MCS aimed to predict the parameters which would result in the most accurate and precise results. The parameters chosen are displayed in Table 5.2.

Table 5.2 Parameters of the ILAP used in the current experiment

Parameter	**Condition N**	**Condition 2K**	**Condition 4K**
Up-down rule	1 down 1 up		
Step size rule (dB)	8,4,2		
Reversals for each step	1,1,6		
Initial Level (IL)	30	40	50

The IL in the Normal condition was set at 30 dBA and was considered acceptable based on the MCS, which found that an IL of 20 and 40 dBA yielded precise and accurate results of SRT50. The reason why the IL was set differently in each condition was because the SRTs obtained in the SCHL were expected to be higher than those from Condition N. According to Levitt (1971), it was advised to set the IL as close as possible to the expected SRT, but when in doubt about the resulting SRT, increasing the IL would not pose an issue as long as the initial step size is large (8 dBA). Therefore, the IL was increased by 10 dB and 20 dB in Conditions 2K and 4K, respectively, to ensure that the IL was audible.

The SRTs obtained from session 2 and 3 were calculated from the mean scored reversal of the last 6 reversals which had a step size of 2 dB. Figure 5.3 is an example of an adaptive procedure of one track (word) for one participant.

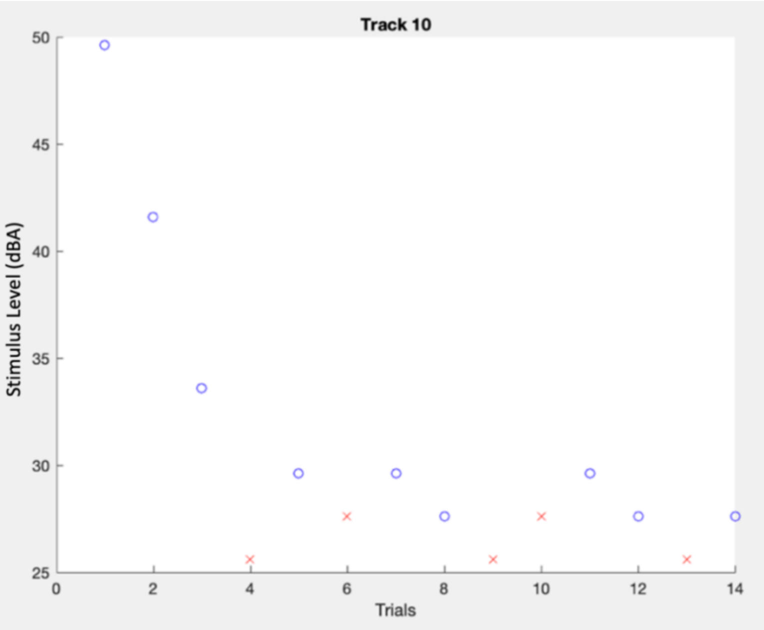


Figure 5.3 A staircase procedure plot of one track (word 10) of participant 19 in condition 4K

5.3.2 Participants

A total of 30 (23 females and 7 males) Arabic speaking, otologically healthy participants aged between 18 and 45 years (mean = 28.3 years) were recruited via email and posters advertising the study. This sample size was based on a previous PhD study that examined the validation of the same test in noise (Al-Kahtani, 2020). All participants were native Arabic speakers living in Saudi Arabia. All participants were Saudi, except for one Palestinian and two Yemenis. The inclusion criteria and recruitment method were like those mentioned in Section 3.2.2.2.

5.3.3 Session 1

Pure-tone audiometry testing, same as the Section 3.2.2.3.

5.3.4 Sessions 2 and 3

The PAAST SiQ was applied through MATLAB software using a code specific for this study. The test consisted of a graphical interface containing 10 words of acoustically similar Arabic names and headphones (Sennhiser HD 650) connected to the computer and placed over the participant's ears.

In Session 2, the participant faced a computer monitor with the interface containing the test words on display (Figure 5.4). The participants were asked to read the words displayed to ensure that they understood them. They were instructed to listen to the sentences ('where is the' + name) through the headphones connected to the computer and click with a mouse on the corresponding name on the screen. They were advised to guess if they did not hear the word or if they were not sure what the word was; they were also encouraged to choose the closest word to what they may have heard. The participant's responses were automatically recorded on the computer.

The participants performed the PAAST SiQ in an ILAP under 3 conditions: Condition N, 2K and 4K, in an entirely random order for all participants. They were allowed a break of 5 minutes (if desired) between each condition. Each condition lasted 11-12 minutes, with most participants requiring 35 minutes to finish the session.

Figure 5.4 PAAST SIQ (after equalisation) graphical interface displayed to participants

Session 3 went on exactly as Session 2. The total test time for all sessions combined was approximately 80 minutes with breaks.

5.3.5 Material, calibration and safety

Material and calibration were similar to section 3.2.2.6.

This experiment was conducted during the COVID-19 Pandemic, shortly after lifting lockdown mandates. The investigator followed Saudi Ministry of Health (MOH) guidance on prevention of contracting COVID-19 (MOH, 2020). Precautions including social distancing, regular handwashing, and face masks for both the investigator and participants were taken. All equipment, tables, and chairs where sanitised using an antibacterial swab/spray between each participant.

5.3.6 Ethical considerations,

Ethical approval (ERGO II: 46958.A2) was granted on 18 November 2020 (Appendix I.4). Ethical approval for the application (569-18) to King Abdulaziz University was also approved on 13 November 2018 (Appendix I.2).

5.4 Analysis Strategy

5.4.1 Introduction

The SRTs of each word in each condition across participants and repeats were obtained through applying the PAAST SIQ in an ILAP and calculating the SRTs. The SRTs obtained from session 2 and 3 were averaged and analysed.

To increase the precision of the results, it is recommended to have at least 6 to 8 recorded reversals before terminating the test (Levitt, 1971). Therefore, it was decided to apply the test twice, and take the average SRTs across repeats, to ensure the precision of the SRTs without risking exhaustion of the participants by increasing the number of reversals, considering there were three conditions. As mentioned earlier the participants needed to complete six reversals in each repeat. Figure 5.5 displays boxplots of the average SRTs (dBA) (across repeat 1 and 2) for each word in the PAAST SiQ across participants in conditions N, 2K and 4K. Throughout the current experiments, the average (of repeat 1 and 2) SRTs were assessed for their sensitivity to SCHL and were referred to as SRTs.

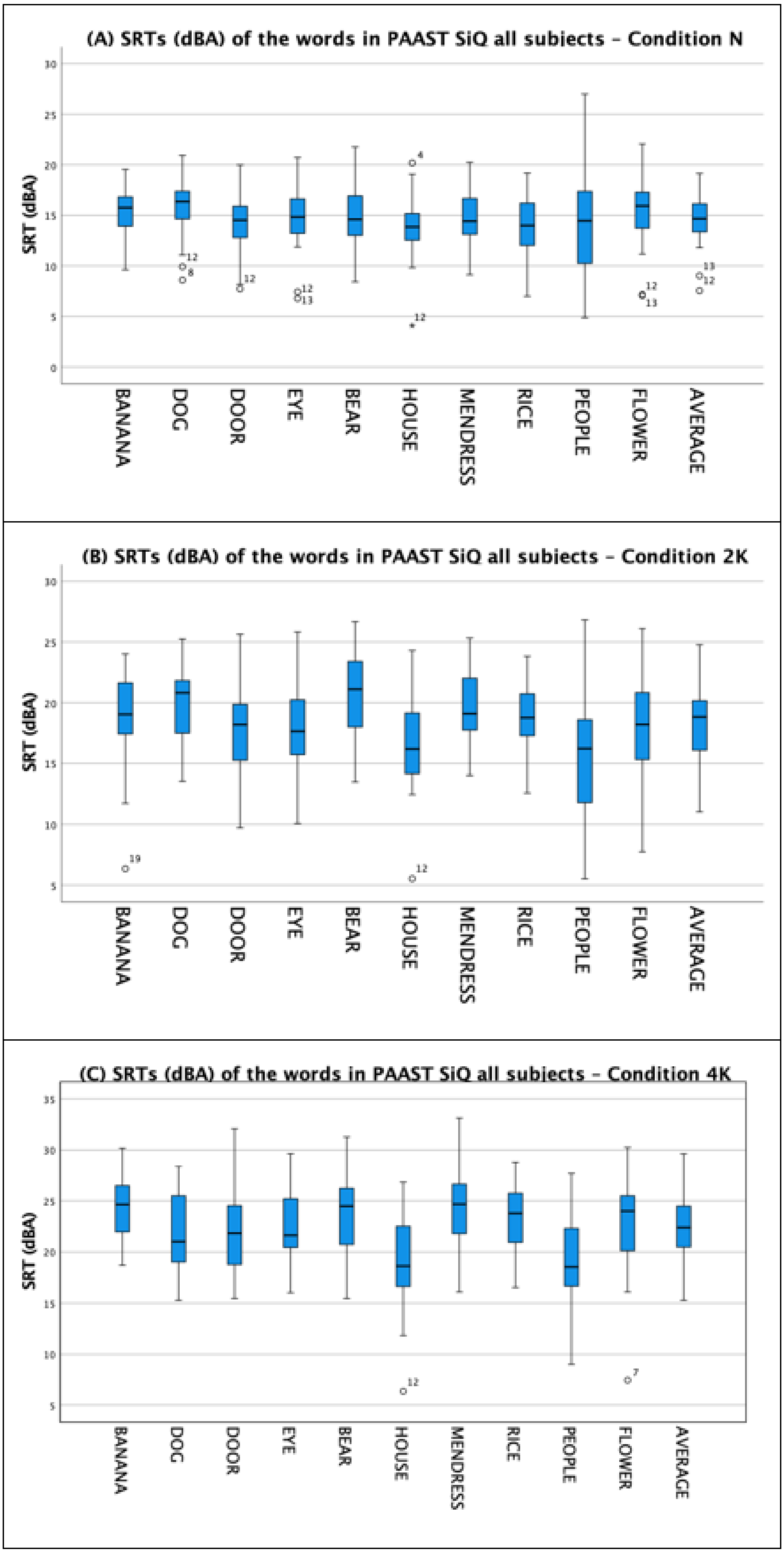


Figure 5.5 Average SRT (dBA) of each word in the PAAST SiQ for all participants across repeats in (A) Condition N, (B) Condition 2K, and (C) Condition 4K.

Descriptive and exploratory analyses of the data were conducted to provide information about the effect of conditions on the SRTs of the words separately and the average SRTs across words, i.e., the average SRT for each participant. Part of the exploratory analysis (as explained in detail at the end of this section) was repeated measure ANOVA (RM ANOVA), which tested the hypothesis as to whether the mean SRTs across words were similar across conditions or not but did not provide information about how similarly or (differently) the SRTs of each word were affected by the conditions. Therefore, obtaining a parameter demonstrating the effect of the change of conditions on each word was necessary.

Predictive analysis can study the behaviour of words across conditions through a regression model, which provides parameters that can be used to compare the effect of conditions between words.

Regression analysis employs a model that describes the relationships between the dependent variables (SRTs) and the independent variables (conditions) (Schneider *et al*., 2010). The resulting linear regression model can be interpreted through several parameters, but for the purpose of analysing the results of this experiment, the focus was on the slope of the regression and the goodness of fit. The slope (b) is a regression coefficient that represented the gradient of the regression line which described the nature of the relationship between the dependent and the independent variables (Field, 2009). In other words, it can describe the change in the mean SRTs associated with change across all conditions. The linear regression model treated the independent variables (conditions) as continuous variables to produce the regression lines, which in turn allowed for estimation of the slope (b) for each participant. A negative relationship of the slope was represented by a falling regression line ($b < 0$), a positive one by a rising regression line ($b > 0$), and no relationship (no change) is represented with a horizontal line ($b = 0$) (Schneider *et al*., 2010). The larger the b value, the steeper the line becomes which represents a bigger change in SRTs across condition as opposed to a smaller b value resulting in a shallow line representing a smaller change in SRTs across conditions. Therefore, words with larger slopes were considered the most affected by the SCHL, whereas words with smaller slope values were less affected by SCHL. The second parameter interpreted was the goodness-of-fit (R^2), which described how well the data fit the model, the closer the value of R^2 to 1, the better the data was fit to the model.

A linear regression model was created for the mean SRTs of each word in the 3 conditions for each participant, below is an example of a linear regression model for the SRTs of each word against the conditions in a participant (Figure 5.6) and the parameters of the regression in the same participant (Table 5.3).

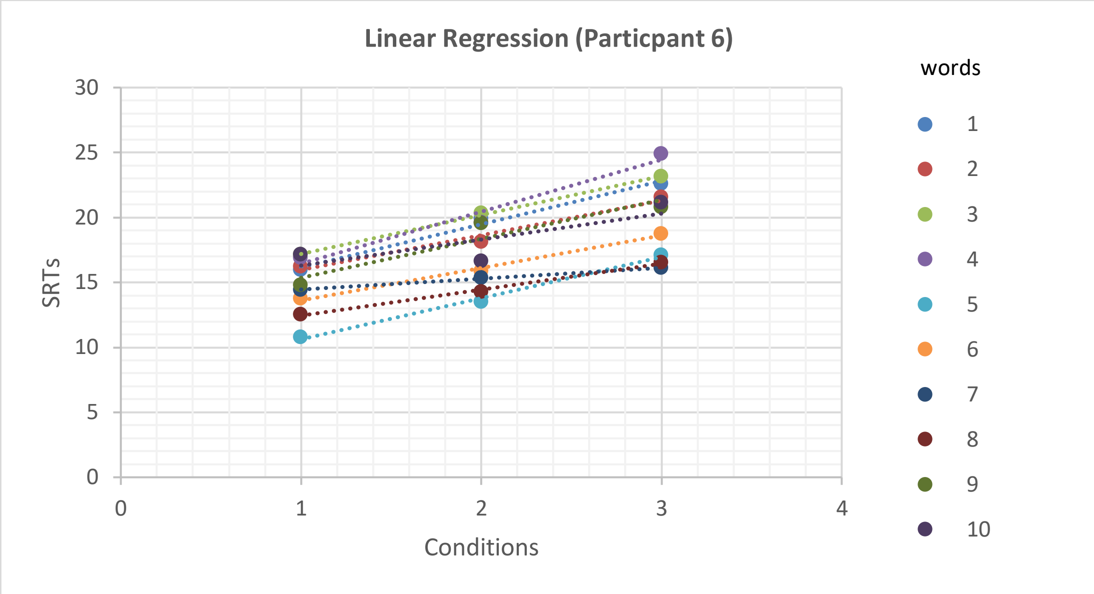


Figure 5.6 Linear regression of the change in each word across conditions in participant 6 (N.B. 1= Condition N, 2= Condition 2K, 3= Condition 4K)

Table 5.3 Parameters of the linear regression model of each word for participant 6

Participant 6							
Word	**Word Number**	**Condition N SRT (dBA)**	**Condition 2K SRT (dBA)**	**Condition 4K SRT (dBA)**	**Average (dBA)**	**Slope**	**R^2**
BANANA	**1**	15.93	19.9	22.6	19.48	3.33	0.99
DOG	**2**	16.23	18.1	21.57	18.63	2.67	0.97
DOOR	**3**	17.13	20.3	23.13	20.19	3	1
EYE	**4**	16.87	19.6	24.87	20.44	4	0.97
BEAR	**5**	10.77	13.5	17.1	13.79	3.17	0.99
HOUSE	**6**	13.73	15.87	18.73	16.11	2.5	0.99
MENDRESS	**7**	14.43	15.33	16.1	15.29	0.83	1
RICE	**8**	12.53	14.27	16.53	14.44	2	0.99
PEOPLE	**9**	14.77	19.53	20.77	18.36	3	0.9
Flower	**10**	17.13	16.63	21.13	18.3	2	0.66

This model was performed for each participant. The resulting slopes and R^2 for each word were averaged across participant, producing one slope and one R^2 for each word.

Additionally, this experiment included the post-equalisation SRTs, which were practically the SRTs obtained from the unattenuated condition (Condition N). This step was considered a follow up to the pre-equalisation stage in Experiment 1.B (Section 4.4.4).

5.4.2 Steps to analysis

- Post-equalisation results
- Descriptive results:
  - Descriptive results of the SRTs
  - Descriptive results of the slopes
- Effect of conditions on the SRTs in the PAAST SiQ
  - RM-ANOVA of the conditions (independent variable, 3 factors) on the average SRT across words (dependent variable) for all participants
  - Independent t-test (against a value of 0) on the slopes of the words
- Effect of conditions on the SRTs of the words
  - RM-ANOVA of the word slopes (10 factors)
  - RM-ANOVA of the words with slope value >4 (4 factors)
  - RM-ANOVA of the words with slope value <4 and >3 (3 factors)
  - RM-ANOVA of the words with slope value <3 (3 factors)

Outliers are identified by measuring residuals, which are the differences between the values predicted by the model and the observed value in the sample (Field, 2017). According to Field (2017), There are three main types of residuals: unstandardized, standardized, and studentized residuals. *Unstandardized residuals* are measured in the same units as the outcome variable therefore, it is not easy to interpret these values across different models. *Standardized residuals* overcome the issue seen with unstandardized residuals by dividing the residuals by an estimate of their SDs. *Studentized residuals* are a variation of standardized residuals, where the unstandardized residual is divided by an estimate of its SD that varies point by point, providing a more precise estimate of the variations in data (Field, 2017). Studentized residuals (SRE) were used in the current study to identify outliers and were calculated for each data point. The Acceptable range of SREs is -3 to +3, and a value outside this range was considered an outlier (Field, 2017).

In analyses that required normality assumption, Shapiro-Wilk test was conducted to determine whether the model residuals could have been produced by a normal distribution. The results of the test of all the variables were assessed based on an alpha value of 0.05, where a $p > 0.05$

indicated that the normality assumption was met. In analysis that require sphericity assumption, Mauchly's test of sphericity was conducted, where a $p > 0.05$ indicated that the sphericity assumption was met (Field, 2017; Mauchly, 1940). In cases where sphericity assumption was violated ($p<0.05$) in RM-ANOVA, the within-subjects factor were calculated using the Greenhouse-Geisser correction to adjust for the violation of the sphericity assumption (Maxwell and Delaney, 2004). The Statistical Analysis was done using SPSS (Version 27).

5.5 Results

5.5.1 Post-equalisation results

The post-equalisation phase was included in the current experiment by measuring the SRTs obtained from testing participants in the Condition N. Observations from the boxplots of the SRTs in the pre-equalisation and post equalisation stage suggested the mean SRT of the words in the post-equalisation stage were very similar to each other (Figure 5.7).

The mean SRTs for all words across the repeats and participants was 14.72 dBA. The range of SRTs was 1.95 dBA. This range was well below the aim range set in the pre-equalisation stage (2.5 dB) indicating these words were very similar in intelligibility (Table 5.4).

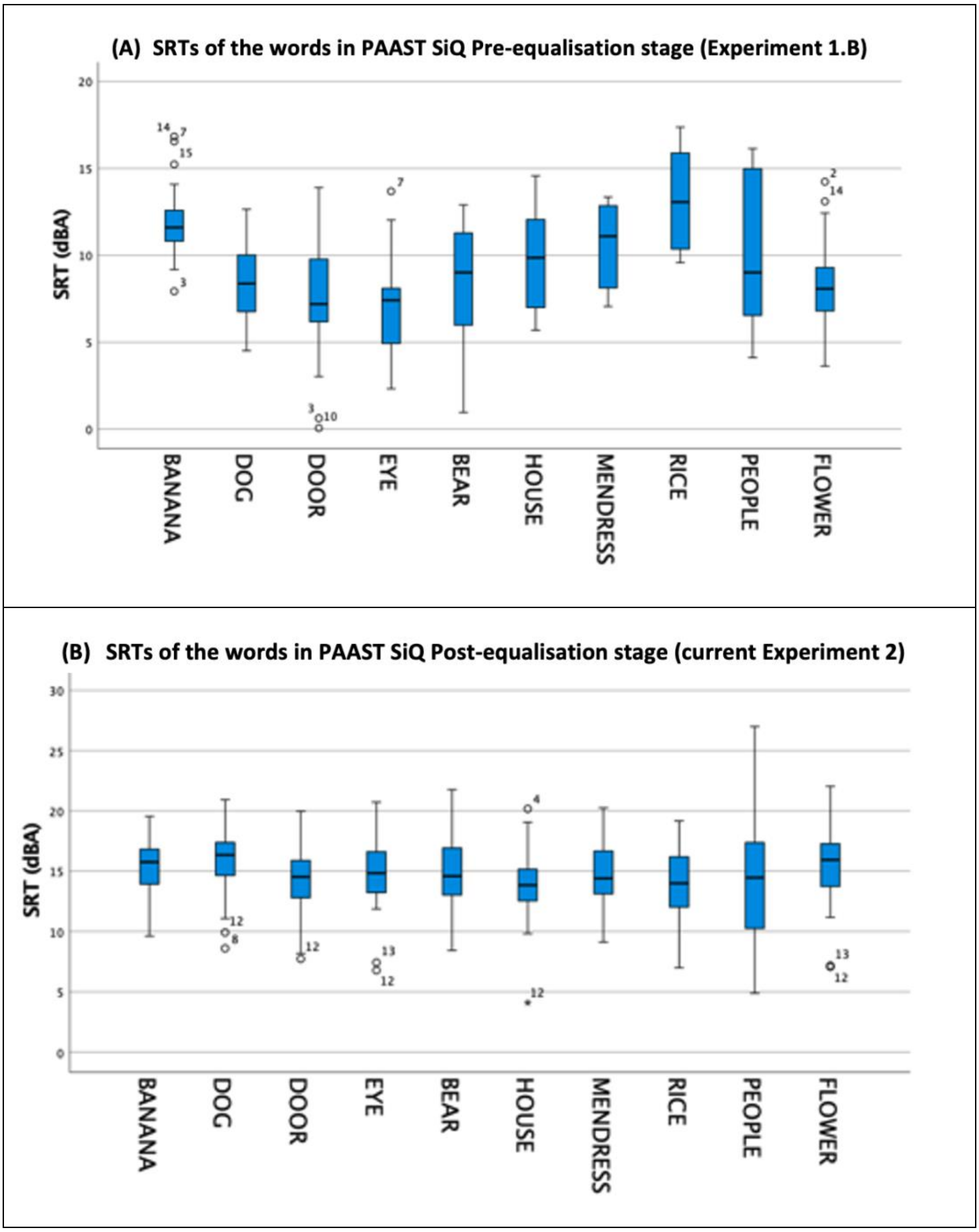


Figure 5.7 Boxplots of SRTs (dBA) of the words of the PAAST SiQ in all participants (A) Pre-equalisation stage (Experiment 1.B) (n = 17), (B) post-equalisation stage (current Experiment 2) (n = 30)

Table 5.4 Mean SRTs (dBA) of each word in Condition N (normal condition) across participants (n=30)

Word	Condition N mean (dBA)	Condition N SD (dB)	Difference from average SRT (dB)
BANANA	15.33	2.68	0.61
DOG	15.77	2.82	1.05
DOOR	14.3	2.79	-0.42
EYE	14.68	3.04	-0.03
BEAR	14.8	3.39	0.08
HOUSE	14	3.22	-0.71
MENDRESS	14.74	2.68	0.03
RICE	13.82	2.98	-0.9
PEOPLE	14.19	5.05	-0.52
FLOWER	15.53	3.47	0.82
Average	14.72	3.21	
Range	1.95		

(Mean = mean SRTs, SD = Standard Deviation from the mean SRT, Diff. from the Average=the difference of each word SRT from the mean SRT of all words.)

5.5.2 Descriptive results

5.5.2.1 Descriptive results of the SRTs

The mean SRTs across participants for each word was averaged across repeats. The SRTs mentioned from now on are those averaged across Repeat 1 and 2.

There was an increase in mean SRTs from Condition N (n=30, mean: 14.72 dBA, SD: 3.21 dB, range=1.95 dB), compared to Condition 2K (n=30, mean: 18.32 dBA, SD: 3.65 dB, range=4.64 dB) and Condition 4K (n=30, mean: 22.37 dBA, SD: 3.88 dB, range=5.70 dB) (Table 5.5) (Figure 5.8).

Table 5.5 Mean SRTs (dBA) of each word in Conditions N, 2K, and 4K

Words	Condition N		Condition 2K		Condition 4K	
	Mean SRT (dBA)	SD (dB)	Mean SRT (dBA)	SD (dB)	Mean SRT (dBA)	SD (dB)
BANANA	15.33	2.68	18.63	4.15	24.58	3.15
DOG	15.77	2.82	19.79	3.36	21.8	3.81
DOOR	14.3	2.79	17.52	3.56	22.02	3.88
EYE	14.68	3.04	17.87	3.62	22.27	3.51
BEAR	14.8	3.39	20.4	3.38	23.67	4.28
HOUSE	14.00	3.22	16.56	3.9	18.99	4.6
MENDRESS	14.74	2.68	19.58	2.92	24.69	3.76
RICE	13.82	2.98	18.94	2.85	23.53	3.1
PEOPLE	14.19	5.05	15.76	4.83	19.33	4.26
FLOWER	15.53	3.47	18.15	3.91	22.86	4.44
Average	14.72	3.21	18.32	3.65	22.37	3.88
Range	1.95		4.64		5.7	

(Condition N = normal condition, no attenuation, 2K = SCHL with a knee point of 2 kHz, 4K = SCHL with a knee point of 4 kHz, *Mean = mean SRTs for each word across participants and repeats, SD = Standard deviation from the mean, Average= average SRTs of all words, range=range of mean SRTs).*

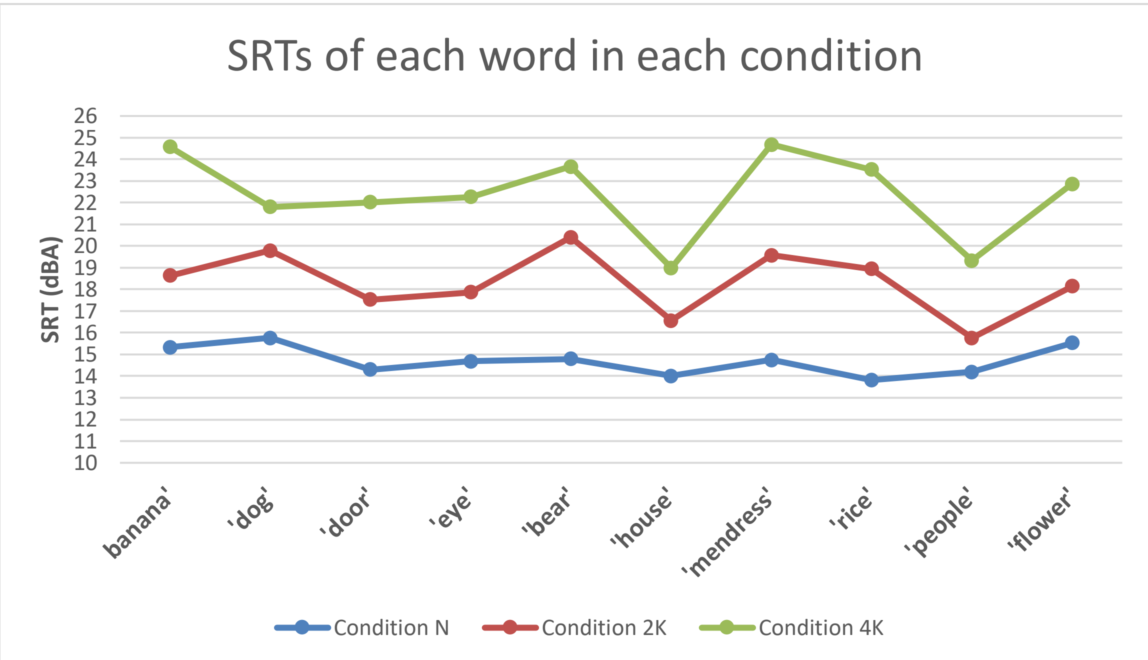


Figure 5.8 Mean SRTs (dBA) for each word across participants and repeats in Conditions N, 2K, and 4K.

As observed from Figure 5.8, although the mean SRTs of each word in Condition N were very similar (range=1.95 dB), the mean SRTs of each word in Conditions 2K and 4K varied. Table 5.6 displays the differences of the SRTs of each word from the mean SRT in the 3 conditions.

Table 5.6 Differences in the mean SRTs of each word between condition 2K and N, 4K and 2K, and 4K and N.

	SRT (2K) - SRT (N) (dBA)	SRT (4K) - SRT (2K) (dBA)	SRT (4K) - SRT (N) (dBA)
BANANA	3.3	5.95	9.25
DOG	4.02	2.01	6.03
DOOR	3.22	4.5	7.72
EYE	3.19	4.4	7.59
BEAR	5.61	3.26	8.87
HOUSE	2.56	2.43	4.99
MENDRESS	4.84	5.11	9.94
RICE	5.13	4.58	9.71
PEOPLE	1.57	3.57	5.14
FLOWER	2.62	4.72	7.33

The word PEOPLE yielded the lowest difference between Condition 2K and N (Diff=1.75 dB) followed by the word HOUSE (Diff=2.56 dB). The smallest difference between Condition 4K and N was seen with HOUSE (Diff = 4.99 dB) followed by PEOPLE (Diff = 5.14 dB). These observations suggested that SCHL may have affected the SRTs of all the words in the PAAST SiQ, but not all words were affected equally by the SCHL (Figure 5.9).

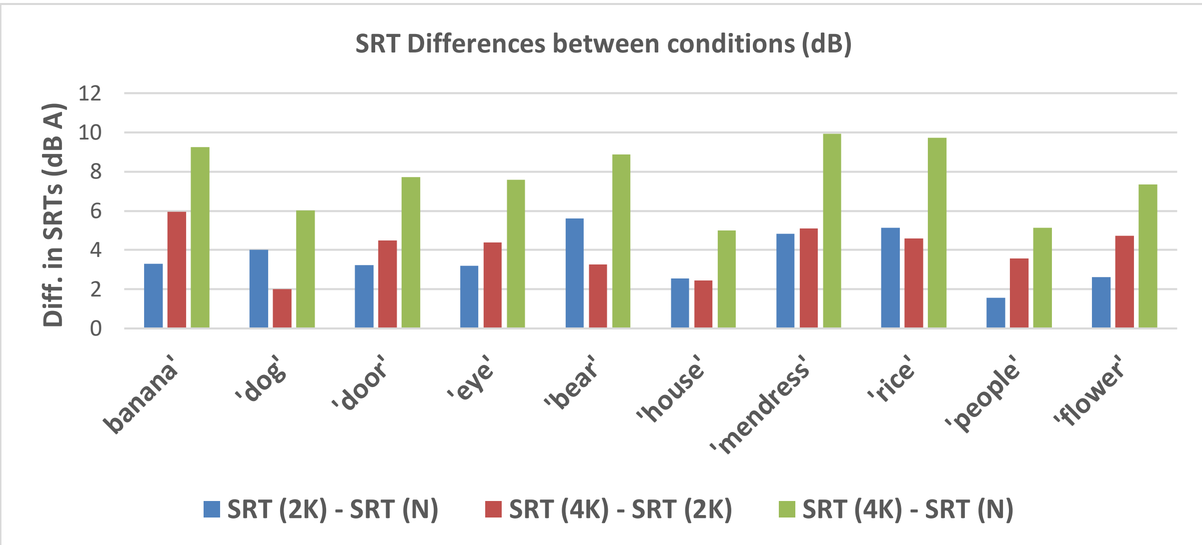


Figure 5.9 Difference in the mean SRT of each word between the three conditions.

5.5.2.2 Descriptive results of the slopes

As mentioned earlier, a linear regression model was fit for the SRTs of each word to the three conditions for each participant. Figure 5.10 displays a linear regression of the mean SRTs of each word average across participants against the three conditions, this served as general representation of how the conditions affect the mean SRTs of the words. The linear regression exhibited a positive change with the conditions, which indicated an increase in SRT from Condition N, to 2K and 4K. The average slopes and goodness-of-fit (R2) of each word SRT (obtained from linear regression of all participants) are displayed in Table 5.7.

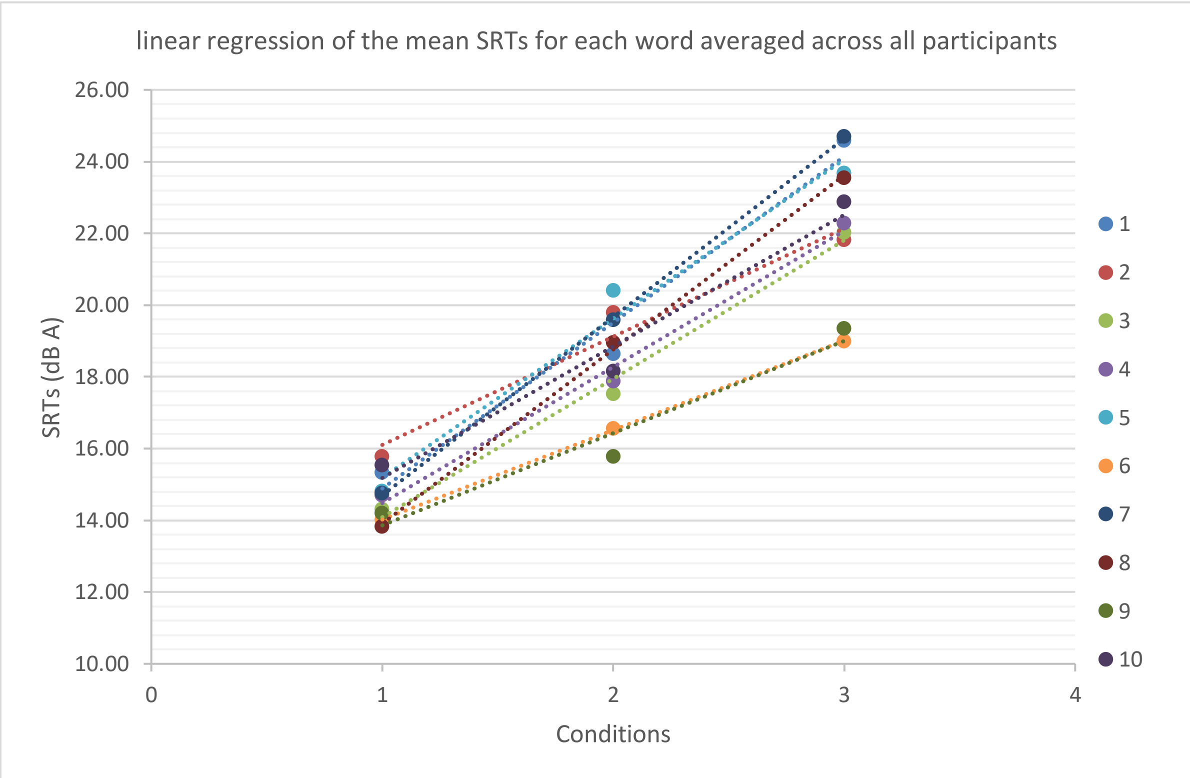


Figure 5.10 Linear Regression of the mean SRTs for each word averaged across all participants against the three conditions (N.B. 1= Condition N, 2= Condition 2K, 3= Condition 4K)

Table 5.7 Parameters of the linear regression of the mean SRTs of the words of all participants

Word	Mean Slope	SD (slope)	95%CI (slope)		R^2
			lower	upper	
Banana	4.63	1.19	4.18	5.07	0.92
Dog	3.02	2.05	2.25	3.78	0.81
Door	3.86	1.14	3.44	4.29	0.93
Eye	3.79	1.19	3.35	4.24	0.94
Bear	4.44	1.93	3.72	5.16	0.93
House	2.50	1.22	2.04	2.95	0.89
Mendress	4.97	1.62	4.37	5.58	0.98
Rice	4.86	1.28	4.38	5.33	0.96
People	2.57	1.93	1.85	3.29	0.73
Flower	3.67	1.62	3.06	4.27	0.87
Average	3.83				0.9

The average slope for all words against the three conditions was 3.83, which indicated a positive change of the mean SRTs with the change of conditions from N to 4K. The words HOUSE and PEOPLE exhibited the smallest slopes (2.49 and 2.57, respectively). The smaller the slope value, the shallower the line was, indicating a smaller change in SRTs of a word from condition N to 4K compared to other words. This information agreed with the descriptive results of the SRTs, where the words HOUSE and PEOPLE yielded the smallest differences from condition N to 4K compared to the other words. This was a preliminary indication that these two words were not as affected by SCHL as the rest of the words. On the other hand, the words MENDRESS and RICE yielded the highest slopes compared to the others (4.97 and 4.86, respectively), resulting in regression lines that were steeper than the other words, indicating that these words were the most affected by the SCHL.

The average R^2, which represented the goodness-of-fit of the line to the SRTs, for all words was 0.9 which can be interpreted as 90% of the variation in the SRTs was due to variation in the conditions. This was generally considered a good fit.

5.5.3 Effect of conditions on the SRTs of PAAST SiQ

5.5.3.1 RM-ANOVA of the conditions across words for all participants

The first statistical analysis aimed to understand how SCHL affected the overall SRTs of the test. The analysis was applied to the mean SRTs of all words for each participant in each condition.

A RM-ANOVA with one within-subjects factor was conducted to determine whether significant differences in the mean SRTs for all words existed among Condition N, Condition 2K, and Condition 4K. Table 5.8 displays the descriptive statistics of the mean SRTs in each condition.

Table 5.8 Descriptive statistics of the mean SRTs for all participants across words

Condition	Mean SRT (dBA)	95%CI (dBA)		Standard Deviation (dB)
		Upper Bound	Lower Bound	
Condition N	14.72	13.74	15.69	2.61
Condition 2K	18.32	17.20	19.45	3.01
Condition 4K	22.37	21.15	23.60	3.27

A boxplot of the mean SRTs (Across words) in each condition (Figure 5.11) indicated that there was an increase in mean SRTs in condition 2K compared to N, and an increase in 4K compared to 2K.

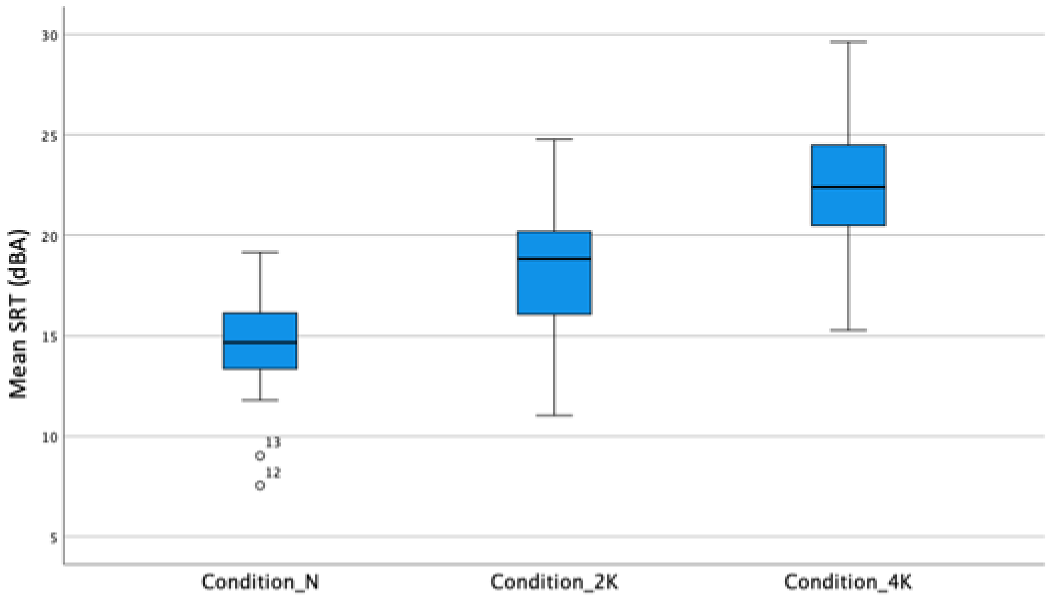


Figure 5.11 Boxplot of the mean SRTs of all words across participants in each condition (n=30)

Studentized residuals (SRE) were calculated for each data point, revealing that all SREs were within the acceptable range (-3 to +3), concluding that there were no outliers in the data (Figure 5.12).

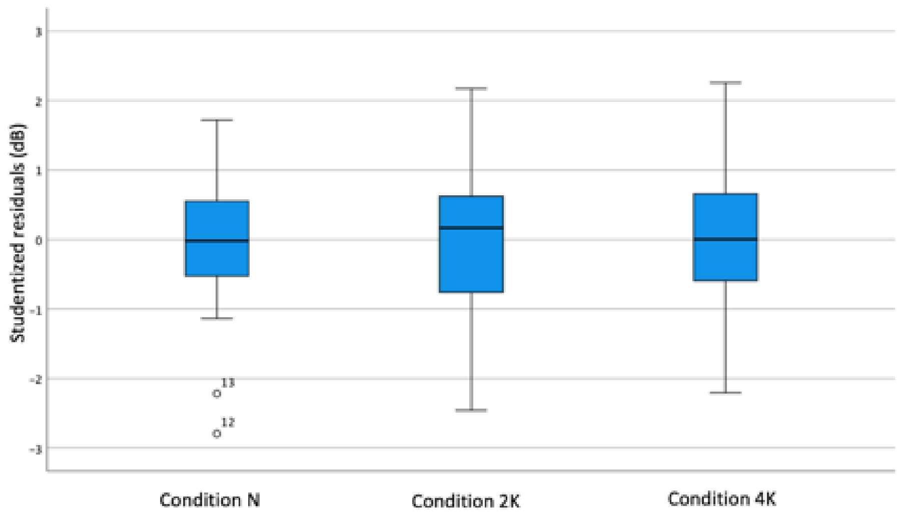


Figure 5.12 Boxplot of the studentized residuals (SRE) (dB) of the SRTs in each condition (n=30)

The normality assumption was met, whereas the sphericity assumption was violated. The main effect for the within-subjects factor (using Greenhouse-Geisser correction) was significant, $F(2, 58) = 317.48$, $p < .001$, indicating there were significant differences between the values of Condition N, Condition 2K, and Condition 4K.

Post-hoc. The mean contrasts utilised Tukey comparisons based on an alpha of 0.05. Tukey comparisons were used to test the differences in the estimated marginal means for each combination of within-subject effects. The reason why Tukey comparison was used was because the sample sizes were equal and population variances were similar (Field, 2009).

Within subject Effects. The mean SRTs (across words and participants) in Condition N were significantly less than those in Condition 2K, $t(29) = -12.67$, $p < .001$, Condition N was significantly less than Condition 4K, $t(29) = -20.24$, $p < .001$, and Condition 2K was significantly less than Condition 4K, $t(29) = -17.55$, $p < .001$. Table 5.9 presents the marginal means contrasts for the Repeated Measures ANOVA.

Table 5.9 The marginal means contrasts for each combination of within-subject variables for the repeated measures ANOVA (Difference (dB) in mean SRTs (dB) across participants

Contrast	**Difference (dB)**	**95%CI**		**P-value**
		Lower	**Upper**	
Condition N – Condition 2K	-3.61	-4.33	-2.88	< 0.001
Condition N – Condition 4K	-7.66	-8.62	-6.70	< 0.001
Condition 2K – Condition 4K	-4.05	-4.63	-3.47	< 0.001

5.5.3.2 Independent sample t-test on the slopes of the words

The descriptive statistics of the slopes of the words indicated that there was a change in SRTs against the Condition from N to 4K. To ensure that the linear regression represented true change in SRT, the null hypothesis, which was that b (slope) = 0 (i.e., no relationship between variables) was tested with a one sample independent t-test. Also calculating the 95%CI can provide information on whether any of the word slopes 95%CI crossed the value 0.

The normality assumption was met for all words. The result of the analysis was significant based on an alpha value of 0.05, with all the word slopes exhibiting a $p < .001$ (Table 5.10), indicating the null hypothesis can be rejected. This finding suggested all word slopes were produced by a distribution with a mean not equal to 0. The 95%CI also indicated that none of the words could have a slope value of 0.

Table 5.10 One-sample t-test of the word slopes against a test value of 0

Word	**Mean Difference**	**95%CI of the Difference**		**p-value**
		Lower	**Upper**	
BANANA	4.63	4.18	5.07	< 0.001
DOG	3.02	2.25	3.78	< 0.001
DOOR	3.86	3.44	4.29	< 0.001
EYE	3.79	3.35	4.24	< 0.001
BEAR	4.44	3.72	5.16	< 0.001
HOUSE	2.50	2.04	2.95	< 0.001
MENDRESS	4.97	4.37	5.58	< 0.001
RICE	4.86	4.38	5.33	< 0.001
people	2.57	1.85	3.29	< 0.001
FLOWER	3.67	3.06	4.27	< 0.001

These results, along with the RM-ANOVA analysis further suggested that the mean SRTs of each word changed (increased) as the conditions change from N to 4K.

5.5.4 Effect of conditions on words

To understand how similarly (or differently) the words are affected by the conditions, a RM-ANOVA with one within-subjects factor was conducted to determine whether significant differences existed among the slopes of the 10 words (10 factors). Normality assumption was met for all words except for FLOWER, therefore the normality assumption was considered to be met. The sphericity assumption was violated.

The main effect for the within-subjects factor of words (using Greenhouse-Geisser correction) was significant, $F(5.437, 157.667) = 16.586$, $p < .001$, indicating there were significant differences between the slope values of the 10 words (Figure 5.13).

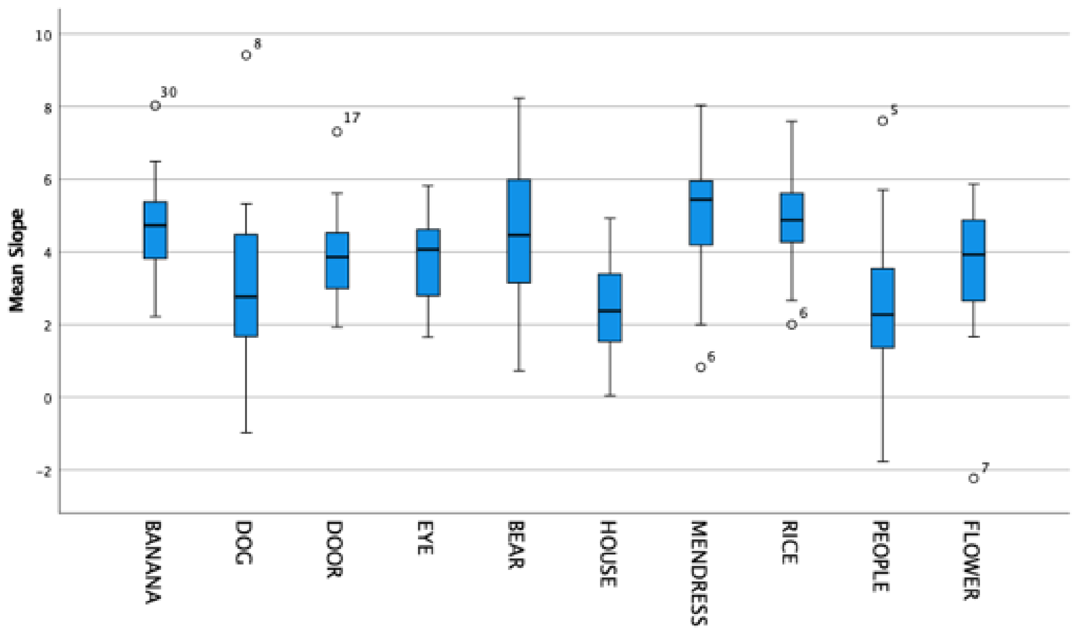


Figure 5.13 Boxplot of the mean slopes for each word for all participants (n = 30).

Considering there were ten words, pairwise comparisons may not have been the most efficient method to determine which words differed significantly. Therefore, classifying the words into groups based on how affected the words were by the change in conditions and analysing each group could give an indication of which words are similarly affected.

5.5.5 Effect of conditions on words in each group

The words of the PAAST SiQ were classified into three groups based on their slopes. A RM-ANOVA with one within-subjects factor (word) was conducted for each group to determine whether significant differences existed among the slopes of the words in each group. The word groups and the result of the RM-ANOVA are displayed in Table 5.11. The results of the RM-ANOVA indicated that the values of the slopes were similar for the words within each group. A boxplot of the slopes of the words in group A, B and C, are displayed in Figure 5.14, Figure 5.15, and Figure 5.16, respectively.

Table 5.11 Groups of words based on their slopes, and the *P*-value result from the RM-ANOVA

Group	**Description**	**Slope value**	**Words**	***P-value***
A	Most affected by change in conditions	Slope value > 4	BANANA – BEAR – MENDRESS - RICE	0.285
B	Moderately affected by change in conditions	4> Slope value >3	DOOR – EYE - FLOWER	0.729
C	least affected by the change in conditions	slope value <3	DOG – HOUSE - PEOPLE	0.294

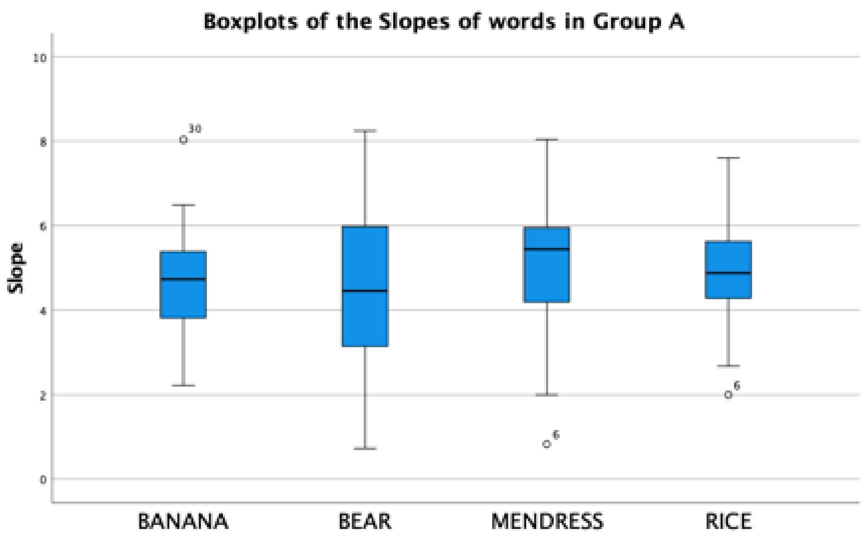


Figure 5.14 Boxplot of the slopes of the words in Group A (4 words, n = 30 participants)

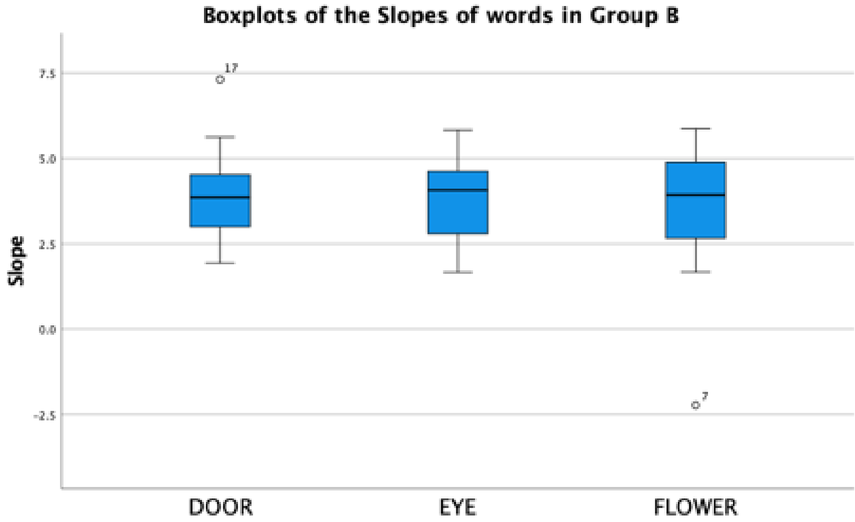


Figure 5.15 Boxplot of the slopes of the words in Group B (4 words, n = 30 participants)

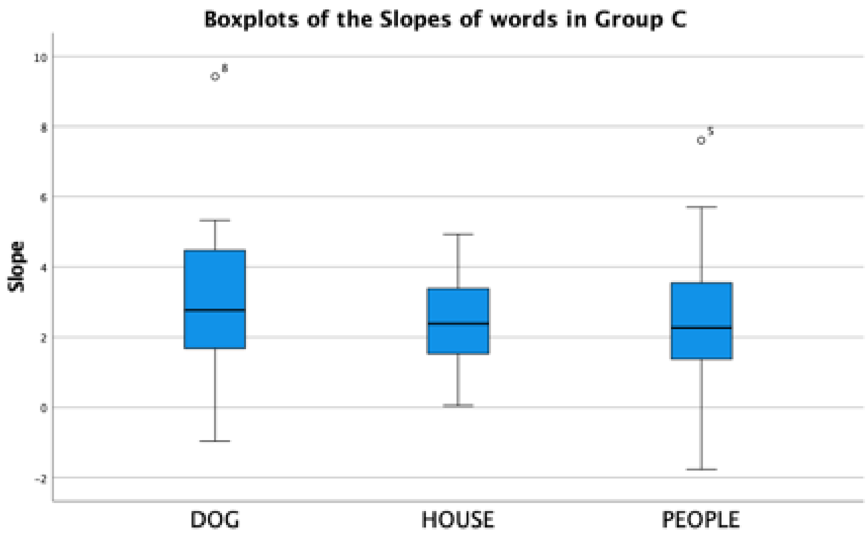


Figure 5.16 Boxplot of the slopes of the words in Group C (3 words, n = 30 participants, outliers represent participants)

5.6 Discussion

5.6.1 Post equalisation of intelligibility of the words of the PAAST

As mentioned earlier in the pre-equalisation stage (Chapter 4), the words of the PAAST SiQ have been equalised for intelligibility by adjusting the RMS of the words by an amount equal to difference of the SRTs of the words from the mean SRT. Also, the decision was taken to discard two pairs (4 words: HEN/ELEPHANT and WORMS/LIGHT) because their required adjustment level exceeded 4 dB, and adjusting by this amount could lead to distortion of the sound of the word (Wagener *et al*, 2003).

The post-equalisation results were obtained from the SRTs in Condition N (no simulation condition) using ILAP. The range of the SRTs was 1.95 dB, which was well below the range set in the pre-equalisation stage (2.5 dB).

Ensuring that the words were similar in intelligibility not only indicated optimisation of the PAAST SiQ, but also can allow to predict that any change in SRTs across conditions would be mostly due to the SCHL.

5.6.2 Effect of SCHL on the PAAST SiQ

The results of this study indicated that SCHL significantly increased the average SRTs across words. There was a significant increase in SRTs ($p < 0.001$) from Condition N (SRT = 14.72 dBA SD = 3.21 dB) to Condition 2K (SRT = 18.32 dBA SD = 3.91 dB), and a significant increase in SRTs ($p < 0.001$) from Condition 2K to Condition 4K (SRT=22.86 dB SD = 4.44). This result suggested the speech recognition ability is worsened with OME-related SCHL.

It is also important to note that the SDs increased from condition N, to 2K and from 2K to 4K. In a study by Moore *et al* (1993), it has also been observed that the SD was somewhat larger in simulated threshold elevation conditions, which was explained by the fact that intelligibility of speech was largely dependent on the more audible component, which in the case of their study, was the low frequency, and that subjects varied in their ability to rely on the low-frequency information to detect the words. This could also be the case in the current study, where individuals varied in their ability to recognise words based on the high-frequency information of these words.

In addition to determining whether there was a significant effect of conditions on the words, it was important to know whether there was a significant change in the SRTs of each word across conditions. As mentioned earlier, linear regression of the SRTs against the condition allowed to demonstrate the change in SRTs across conditions using slopes. Each of the word slopes had a positive value significantly larger than 0, indicating that there was a positive change in the SRTs of each word across conditions from Condition N to Condition 4K, with an average slope of 3.83. Since there hasn’t been a study with similar objectives and procedures, it would be hard to judge whether the values of the average of the slope indicates a strong change, but in general, a slope of 3-5 was considered a gentle slope (Slope Steepness Index, 2021).

Several studies attempted to measure the effect of SCHL resembling that of OME on speech recognition. A study by Cai *et al.* (2017) measured the reception thresholds for sentences (RTS) obtained through the Mandarin hearing in noise test in both noise and in quiet. They found that the results of RTS in quiet were similar in children with actual OME and otologically healthy children who performed the test in HLS conditions, whereas the RTS in noise was significantly better in children with actual OME. This finding ensured that threshold attenuation in SCHL can resemble that of actual CHL associated with OME. Also, it indicated that measuring speech in quiet in CHL is a better predictor of the degree of hearing loss compared to speech in noise based on the hearing loss framework (Plomp, 1978) explained in literature review.

Another study by Penn *et al*. (2004) studied the effect of SCHL on speech recognition scores (% correct) in adults and children, using Balanced Kindergarten (PBK) and Nonsense Syllable Test word lists. The conditions applied were unattenuated, average attenuation, and maximum attenuation. The SCHL conditions (both average and maximum) had more attenuation (worse thresholds) in the low and high frequencies than in the 2kHz frequency. Penn *et al* (2004) found a significant effect of SCHL conditions ($p < 0.001$) and age ($p <0.001$) on the speech scores. The results of the study resembled the results of the current study. The worse the simulated condition was, the worse the outcome of the speech test. The fact that there was a significant effect of attenuation on speech recognition in children in the study by Penn *et al.* (2004) predicted that the PAAST SiQ could be affected by SCHL associated with OME in children as well. Another important point to address was the configuration of simulation in the study by Penn *et al* (2004), where they attenuated both the low and high frequencies, and even the 2kHz frequency in the maximum attenuation condition. The current study looked at the effect of simulated low frequency CHL alone, sparing the 2 kHz and the 4 kHz in conditions 2K and 4K, respectively. Given that low frequency CHL is the most common audiometric configuration in children with OME (Cai *et al.*, 2018), it was important to ensure that the PAAST SiQ would be able to detect hearing loss in these frequencies and not be affected by normal higher frequency thresholds.

The significant difference in SRTs between conditions indicated that, in general, the words of the PAAST SiQ are sensitive to low frequency hearing loss, leading to the second aim of the current study, which was to look at whether the words of the PAAST SiQ were equally sensitive to SCHL.

5.6.3 Effect of SCHL on the words of the PAAST SiQ

The purpose of achieving the aim of investigating whether the words were equally sensitive to SCHL arose from the fact that, as mentioned earlier in the introduction, the MTT can be applied with as few as 3 pairs of toys/words (Hall *et al*., 2007; Lovett *et al*., 2013). In their study, Hall *et al*.

(2007) found no significant difference in word recogntion based on the number of the toy pairs used (p > 0.05). In her thesis, Al-Kahtani (2020) applied the PAAST SiN to children, using seven pairs in older children and five pairs in younger children unfamiliar with all words. In all these studies, it was not specified which word pairs were excluded. The reason why was probably because like the English version (MTT), the PAAST was composed of monosyllabic nouns that were familiar to children, representing all manners and place categories of the language (in case of the PAAST, the Arabic language) consonants. It was important for the child to be familiar with all the words to perform the test, but understanding how each word would behave in a SCHL addresses the gap in knowledge as to whether we could comfortably eliminate words from the PAAST SiQ without risking the under-detection of CHL.

Repeated Measure-ANOVA of the ten words slopes revealed that there was a significant main effect ($p < 0.001$) of the words on the slopes, indicating that there was a significant difference in the effect of SCHL between the words. The words MENDRESS and BANANA had the largest slope value (4.97 and 4.63, respectively), which meant these words were the most affected by the change in conditions. On the other hand, the words HOUSE and PEOPLE were the least affected by the change in conditions (slope value 2.49 and 2.57, respectively), indicating that the SRTs of these words did not increase as much as the other words.

To have a better understanding of the effect of conditions on the words, the words were grouped based on their slopes. The decision to do so arose from the observation that there was a trend among words where some share similar slopes with other specific words, indicating that the words within each group may behave similarly to SCHL.

The first group of words (Group A) had a slope value of >4. These words were BANANA, BEAR, MENDRESS, and RICE. Analysis revealed there was no significant difference between the slopes of these words (p = 0. 037), suggesting that there was no significant difference in how the SRTs of words in this group increase with SCHL. This group of words exhibited the largest effect in their SRTs with the change of conditions. This could be explained by the high frequency component of the words BANANA, BEAR, and RICE. The word MENDRESS did not have as much high frequency information as the other words in the group, but its high intelligibility could be explained by the long duration of the vowel (/ɔ/).

The second group of words (Group B) contained words with a slope value <4 and >3. The words in this group were DOOR, EYE, FLOWER. Analysis revealed there was no significant difference between the slopes of these words (p = 0.578). This group of words exhibited a moderate effect in their SRTs with the change of conditions compared to the words in Group A.

The third group of words (Group C) contained words with a slope value <3. The words in this group were DOG, HOUSE, PEOPLE. This group of words exhibited the least effect in their SRTs with the change of conditions compared to Group A and B. The result of Group C was unexpected because the word PEOPLE contained more high frequency information (consonant /s/) than some words in group A and B, but this might be explained by the low amplitude of the consonant at the end of the word. It can be observed that two pairs of words in the PAAST (pair 1 and 2) are made up of the words in group A (word 1 and 2), and pairs 3, 4, and 5 consisted of the words from Group B (word 1) and the words from group C (word 2) (Table 5.12).

Table 5.12 Pairs of the words of the PAAST SIQ and the groups each word belong to

Pair	word 1	word 2
1	BANANA (Group A)	MENDRESS (Group A)
2	BEAR (Group A)	RICE (Group A)
3	DOOR (Group B)	PEOPLE (Group C)
4	EYE (Group B)	HOUSE (Group C)
5	FLOWER (Group B)	DOG (Group C)

This finding suggested that word pairs 1 and 2 ideally should not be eliminated from the PAAST SIQ when testing children, because their elimination could result in not identifying sight/mild hearing loss. on the other hand, the elimination of either one of pairs 3 ,4, or 5 may not significantly affect the sensitivity of the PAAST SiQ to CHL associated with OME.

5.7 Conclusion and Recommendations

- The words of the PAAST SiQ (Post-equalisation) are considered equal in intelligibility
- The PAAST SiQ is sensitive to SCHL, the average SRTs across words increased significantly from Condition N to Condition 4k.
- The mean SRT of each word in the PAAST SiQ increased significantly with SCHL.
- The increase in SRTs across condition was not equal for all words in the PAAST SiQ, meaning the words did not behave similarly to SCHL.
- The words were grouped into 3 groups (A, B, C), where group A (word pairs: BANANA – MENDRESS and BEAR – RICE) exhibited the largest change in SRTs across conditions indicating that elimination of words from this group may affect the ability of the PAAST SiQ to detect slight/mild hearing loss. This suggestion can be confirmed by testing children with OME with different CHL severity levels and assessing the ability of the PAAST to detect slight CHL using less pairs of words.

Chapter 6 Experiment 3: Developing an Arabic quality-of-life questionnaire (AOM6) for children with Otitis media

6.1 Introduction

6.1.1 Background

The aim of this study was to develop an Arabic questionnaire to assess QoL in children with OME, by adapting a validated QoL questionnaire. Arabic is the official language in 27 countries spread between North Africa and the Middle east, and currently there are no published studies on questionnaires that assess the QoL in children with OME in the Arabic language. The OM6 was chosen to be translated and cross-culturally adapted to the Arabic language because it is a simple, valid, and reliable questionnaire, it is also the most commonly used and translated disease specific QoL questionnaire (Timmerman *et al.*, 2007; Gan *et al.*, 2018). The OM6 was designed to assess the QoL in children with otitis media (OM), which can be divided into two major diagnostic subgroups: Acute otitis media (AOM) and otitis media with effusion (OME) with great overlap between the two (Heidemann *et al.*, 2013; Tao, Schulz, Donna B. Jeffe, *et al.*, 2018). The subgroups of OM are mentioned in Table 6.1.

Table 6.1 Definitions of diagnostic subgroups of otitis media (adapted from Heidemann *et al*., (2013) and Tao *et al*., (2018))

Subgroup	**Definition**
Acute Otitis Media (AOM)	Middle ear effusion and acute onset of signs and symptoms of middle ear inflammation such as fever, otalgia, possible otorrhoea and discomfort that may result in interference with or precludes normal activity or sleep.
	Recurrent acute otitis media (rAOM) is defined by the presence of at least 3 episodes of AOM in 6 months or 4 or more episodes in 1 year.
Otitis Media with Effusion (OME) ***or*** **Chronic Otitis media (COM)**	Middle ear effusion without signs or symptoms of acute ear infection. Disease severity of OME ranges from no symptoms to lowered activity level and sleep disturbances or even significant hearing loss and speech impairment.

The OM6 questionnaire consists of a Functional health status (FHS) part composed of six items: physical suffering, HL, speech impairment, emotional distress, activity limitations, and caregiver concerns, and a global QoL in child assessment using a visual analogue scale (VAS). The OM6 is a proxy-questionnaire, meaning that it is filled out by the caregiver, who is most often a parent. There are seven response options to each FHS item: not present/no problem, hardly a problem at all, somewhat of a problem, moderate problem, quite a bit of a problem, very much of a problem and extreme problem. A mean score of all six items comprises the OM6 Total Score. The highest score would be a score of 7, and the lowest score would be 1 (Timmerman *et al.*, 2007). Although there is no normative data set for the OM6, high scores generally indicate a poor HR-QoL (Gan *et al*., 2018). The VAS is a 10-Likert scale response option, with illustrations of facial expressions at each scale point. A low score indicated a bad global QoL, and high score indicated a good global QoL. The scores of the VAS serve the purpose of inquiring about the global QoL in children with OME and can be used to assess construct validity but are not included in the OM6 score. Details about the rationale of using OM6 and its validation studies are mentioned in Section 2.3.2.

It could be considered sufficient to translate a well-validated questionnaire such as the OM6, which has been validated and tested for its psychometric properties in English and other languages (Rosenfeld *et al.*,1997; Brouwer *et al.*, 2005; Heidemann *et al.*, 2013; Tao *et al.*, 2018). Ideally, it is recommended that further steps be taken to examine the psychometric properties, to ensure that the translated version serves the same aim as the original one (Hall *et al*., 2018). Examining the construct validity, reliability and responsiveness of the translated questionnaire can be laborious and time-consuming work, but it ensures that the translated questionnaire conveys the same measures as the original. Assessment of the clarity and cultural acceptance through field testing, can provide some measure of quality control for the content. (Hall *et al*., 2018).

6.1.2 Translation, cross-cultural adaptation, and psychometric properties testing of quality-of-life questionnaires

Health-related QoL assessment has markedly evolved in the last two decades, mainly in the English language. International interest in HR-QoL assessment has also increased due to the inclusion of QoL as an outcome measure in clinical trials, an important facilitator of cross-cultural study comparisons. Therefore, the demand to develop QoL assessment tools by translating validated questionnaires into different languages has increased over the years, including in the Arab world. Spoken Arabic differs between countries (as well as regions within the same country) because of the different dialects of each area. At the same time, Literary Arabic (used in most written documents, medical assessments, and formal spoken occasions) is the same across all

Arabic-speaking countries. One of the earliest reports of translated QoL questionnaires to the Arabic Language was in 1998 (Al Sayah *et al.*, 2013). Since then, these measures' cross-cultural adaptation and validation in the Arab world have increased. Cross-cultural adaptation includes both translation and cultural adaption to the population that will use the questionnaire (Hall *et al.*, 2018).

The main goal for cross-culturally adapting the OM6 was to apply it to an Arabic-speaking population, not only so that they can understand the questionnaire but also to achieve equivalence between the original and the translated version of the scale. Several types of equivalencies have been proposed by Streiner *et al.*, (2015), where they focused on five key equivalencies (Table 6.2) that were adapted in the translation of the OM6 questionnaire to ensure that the translation is not 'word for word' but is instead a 'word for world' translation. The first four types of equivalencies (conceptual, item, semantic, and operational equivalencies) can be achieved through knowledge of the target group's culture from ethnographic literature reviews, interviews, and consultations with a group of experts.

Table 6.2 Key types of equivalencies for translating a quality-of-life questionnaire (Adapted from Streiner *et al*. (2015))

Equivalency Type	Details	Example
Conceptual equivalence	Both cultures agree on the elements that constitute the construct. The opposite of conceptual equivalence is that the concept may not exist in the target culture.	A questionnaire that measures how much a person enjoys their life applied to a person living in poverty and does not have an existential relevance of enjoyment.
Item equivalence	Determines whether the specific items are relevant and acceptable in the target population.	Inquiring about sexual problems or negative feelings toward family in a culture that considers these issues taboo.
Semantic equivalence	Refers to the meaning attached to each item.	The phrase 'I feel blue' would not convey the same semantic meaning in another culture.
Operational equivalence	Looks at whether the same format of the scale, the instructions and the mode of administration can be used in the target population	In some cultures, it is impolite to ask direct questions, and in other cultures it is impolite for young people to direct questions to their elders.
Measurement equivalence	Ensuring the psychometric functions are the same in both versions	To make sure both versions are reliable and valid through testing the translated version of the questionnaire.

The last type (measurement equivalency) is achieved by testing the psychometric properties of the questionnaire and comparing them to the original version. This was partly achieved in stage II, where psychometric properties were assessed, including internal consistency, construct validity and floor and ceiling effects.

6.2 Aim and objectives

The aim of this study was to develop an Arabic questionnaire to assess the QoL of children with OME, by adapting it from an existing valid QoL questionnaire, and to ensure that the Arabic version is culturally acceptable, easy to understand and is a valid questionnaire. This aim was set to be achieved in three stages:

Stage I Translation

To translate the OM6 questionnaire to the Arabic language according to the guidelines for translating HR-QoL questionnaires and ensure equivalence to the original OM6 through a committee review. This stage resulted in the Arabic version of the OM6: AOM6.

Stage II Assessment of the psychometric properties of the AOM6

Apply the AOM6 to parents of children diagnosed with OME/rAOM, to assess the scores of the questionnaire and assess psychometric properties of the AOM6.

Stage III Assessment of the clarity and cultural acceptance of the AOM6

Ask parents who participated in stage II to complete a field test questionnaire to assess the clarity and cultural acceptance of the AOM6

6.3 Stage I: Translation and cultural adaptation

The translation and cultural adaptation process was adapted from two main guidelines (Beaton *et al*., 1999; Hall *et al*., 2018) (Table 6.3). A literature review was conducted to search for Arabic questionnaires that assessed QoL of children with OME. No Arabic version of the OM6 was found, nor was there any other questionnaires that measure QoL of OME in children. The OM6 was chosen to be translated and adapted in Arabic because of its brevity, ease of scoring, and good psychometric properties. The developer of the OM6, Dr. Richard M. Rosenfeld, was contacted via email to explain the research idea and obtain permission to translate the OM6 into Arabic. Permission to translate the OM6 was granted on 25th December 2018, and they also confirmed that no Arabic OM6 questionnaire was available, to their knowledge, to date (Permission attached in Appendix section G.1). Two forward translations from English to Arabic by two different translators were conducted. The researcher chose a single translation from the two forward translations. The chosen translation had slightly better semantic equivalence. Changes to some words were applied, resulting in Version 1 (V.1) of the questionnaire. One backward translation of Version 1 from Arabic to English was completed by a third translator (T3), which was almost identical to the original English version. Afterwards, a committee review was held consisting of the researcher, an Audiovestibular Medicine consultant, and a translator to discuss Version 1 of the questionnaire. Changes included choosing terms that were commonly used to describe ‘delayed speech’ and ‘poor pronunciation’ in Arabic. These terms, when translated word for word, gave an exaggerated, negative meaning. Therefore, the terms were changed to the more appropriate 'تأخر في الكلام' and 'ضعف في النطق', respectively. The committee review also agreed that the Arabic OM6 was easy to understand, culturally acceptable, and satisfied all equivalency key points mentioned in Table 6.2, except the measurement equivalency, which required further testing of the psychometric properties.

Table 6.3 Summary of the process of translation of the OM6 (adapted from Beaton *et al.*, (1999) and Hall *et al.*, (2018))

Stage	Section	Item	Description
Translation	***I preparation***	1	Identified whether a target language version exists
		2	Permission from the developer (Richard M. Rosenfeld)
	II Translating the source language into the target language	1	First forward translation by Translator 1
		2	Second forward translation by Translator 2
		3	Single reconciled translation produced from the two translations by Translation Coordinator
	III Translating the target language back into the source language	1	Backward translation by Translator 3
	IV committee review	1	1 translator + 1 clinician + Translation Coordinator: cross-cultural adaptation (Changes were applied to translation)

After the translation process was complete, an Arabic version of the OM6 (V.2) was established and was ready to be tested for its psychometric properties and cultural acceptance. The English and Arabic versions of OM6 questionnaires are attached in Appendix G.3.

6.4 Stage II: Measurement of the psychometric properties of the AOM6

6.4.1 Aim and objectives

To assess aspects of reliability and validity of the AOM6. This was achieved by obtaining responses on the AOM6 from parents of Arabic speaking children with OME/rAOM. The main objectives were:

1- To obtain and describe the scores of the AOM6.
2- To assess the following psychometric properties: Internal consistency, construct validity, and floor and ceiling effect.

6.4.2 Method

Participants

The decision of the sample size was based on guidelines on assessing the psychometric properties of a questionnaire (Al Sayah *et al.*, 2013; Tao *et al.*, 2018), which suggested that an adequate sample size would be a number equal to 7 times the number of items, which in the case of AOM6 was 6 items, resulting in n = 42, but at the same time it would be preferable to have a number above n = 100. Given the scope of this PhD, it has been decided that 50 participants would be sufficient for this study, using convenience sampling. The inclusion and exclusion criteria are mentioned in Table 6.4.

Table 6.4 Inclusion and exclusion criteria for participants in Stage II of the development of the AOM6

Inclusion	- Parents/caregivers of children aged between 6 months and 12 years - Native Arabic Speakers - Children diagnosed with o rAOM (Recurrent Acute Otitis Media): ≥ 3 episodes of Acute Otitis Media in the past 12 months with full recovery in between episodes o OME/COM: Presence of effusion in the middle ear ≥ 3 months - No surgical treatments for the otitis media in the past 3 months
Exclusion	- Parents/caregivers of children with: o Syndrome diseases e.g cleft palate o Other concurrent diseases that can affect the quality-of-life e.g cardiac or respiratory disease o Children diagnosed with permanent SNHL or CHL

Parents of children fitting the inclusion criteria were approached in the ENT clinic in King Abdulaziz University and InterMed clinic, Jeddah, Saudi Arabia. The diagnosis was made by the ENT doctor in the aforementioned centres.

Data were collected between August and November 2021. Parents of children who fit the criteria were approached and the study was explained to them. Parents who agreed to fill the questionnaire were 42 parents, all were Saudi except for 2 Egyptians and 3 Yeminis (same nationality as their children). This study was mostly conducted in King Abdulaziz University hospital, but additional data was collected from a private clinic (InterMed clinic).

This current validation study included only children with OME or rAOM, because the OM6, which the AOM6 was adapted from, is a disease specific QoL questionnaire, therefore data from normal children was not included.

Study design and procedure

Parents of children fitting the inclusion criteria were approached, and the study was explained to them. Upon agreement to participate, they were asked to sign the consent forms. They were then asked to fill a form (attached in Appendix G.4) that contained the following:

1- **Demographic data of the child:** Date of birth, sex, and an ID number given by the researcher.
2- **Information about the ear disease** *(this was filled with help from the researcher based on the diagnosis made by ENT which was stated in their notes)*: Type of infection (rAOM or OME), side of infection, and the inquiry about other chronic diseases.
3- **The AOM6 questionnaire**
4- **Field-testing:** Details are mentioned in 6.5.2.

The questionnaires were collected and transcribed in an Excel sheet for analysis.

Ethics and Safety

Ethical approval (ERGOII 52801.A2) for this experiment was granted from the University of Southampton on 03 August 2021 (see Appendix I.5). Ethical approval for the application (755-19) from King Abdulaziz University was also approved on December 31, 2019 (see Appendix I.6). This experiment was conducted during the COVID-19 Pandemic, shortly after lifting lockdown mandates. The investigator followed Saudi Ministry of Health (MOH) guidance on prevention of contracting COVID-19 (MOH, 2020). Precautions including social distancing, regular handwashing, and face masks for both the investigator and participants were taken. All equipment, pens, tables, and chairs where sanitised using an antibacterial swab/spray between each participant.

6.4.3 Analysis strategy

Descriptive results

The method of scoring of the AOM6 was adapted from previous studies that aimed to validate the OM6 (Rosenfeld *et al.*,1997; Lameiras *et al.*, 2017). The mean score of each question was calculated from the 7-point Likert scale scores across participants. The frequency of the responses to each scale point for each question was also calculated.

The total score of the AOM6 (AOM6 Total) was calculated as the sum of the scores of the 6 items divided by the number of items (6). As later explained in the section on **Internal consistency**, there are two subscales in the AOM6, the AOM6 HaS and AOM6 BaS. The AOM6 HaS score was calculated as the sum score of items 2 and 3 divided by the number of items (2), and AOM6 BaS was calculated as the sum score of items 1, 4, 5, and 6 divided by the number of items (4).

The AOM6 total and subscale scores were also analysed based on presence of unilateral or bilateral otitis media.

To assess whether the AOM6 can differentiate between rAOM and OME, an independent t-test was carried out to measure the difference in AOM6 total score in between rAOM and OME.

Floor and ceiling effect

Floor and ceiling effects were considered present if more than 15% of participants achieved the highest or the lowest possible score (Heidemann *et al*., 2013). According to Streiner *et al*., (2015) The presences of floor and ceiling effects would suggest that items in the questionnaire might not truly represent QoL and that some items are being wasted. It would then be important to modify the items so that most scores lie in the middle of the scale, effectively reducing the variance in performance (Streiner *et al*., 2015).

Internal consistency

The internal consistency measures an aspect of the reliability of the questionnaire. The underlying factor structure of the AOM6 is adapted from an OM6 validation study (Tao *et al.*, 2018), in which they used principal components analysis (PCA) with varimax rotation to determine two factors: Behaviour and Symptoms (BaS) and Hearing and Speech (HaS).

In the study by Tao *et al.*, items measuring Physical Suffering, Emotional Distress, Activity Limitations and Caregiver Concerns were highly correlated with each other and loaded onto BaS. Items measuring HL and Speech Impairment were highly correlated with each other and loaded onto HaS.

These factors were used for this study to assess the internal consistency of BaS, HaS, and the Total AOM6 using Cronbach's α coefficient. Streiner *et al.* (2015) recommend that for a QoL questionnaire, Cronbach's α should be above 0.70 to be considered to have a good internal consistency.

Construct validity

Construct validity is a framework of hypothesis testing based on the knowledge of the underlying construct (Streiner *et al*., 2015), which in the case of the current study is OM. In OM6, for example, a hypothesis that the number of days the child had decreased activity would correlate positively and strongly with the activity limitation (Q5) item (Heidemann *et al.*, 2013). If the theory was correct, and the test was valid, then the correlation would be as hypothesized (Streiner *et al*., 2015). There could be an endless number of hypotheses that can be constructed based on prior knowledge of the disease, and the general rule is that at least 75% of the results are in accordance with these hypotheses (Timmerman *et al.*, 2007).

A hypothesis was constructed regarding the correlation between the AOM6 and the Visual Analogue Scale (VAS) based on findings from previous studies (Rosenfeld *et al.*, 1997; Heidemann *et al.*, 2013). Ideally, several hypotheses should be constructed, based on the possible correlations between the tool being tested (in the case of this experiment AOM6) and the results of other tools or findings, but the scope of the study allowed for assessing a single hypothesis on the relationship between the AOM6 and the global QoL of the child. Negative significant correlation was expected between the AOM6 Total and the VAS score of AOM6, because it is likely that the global QoL of life assessed by VAS would be low (worse) if the AOM6 scores were high (worse). A correlation of <0.3 was defined as weak, 0.3-0.5 as moderate and >0.5 as strong (Heidemann *et al.*, 2013).

All statistical analysis were performed using SPSS 28.0.1.1.

6.4.4 Results

As mentioned before this experiment was conducted during the COVID-19 pandemic, around the time lockdown mandates were starting to gradually be lifted. Patient flow was still affected because parents were worried about taking their children to hospitals, which resulted in recruiting fewer participants (n = 42) than the proposed sample size (n = 50). It has also been noticed that the number of children with rAOM/OME decreased during the pandemic compared to previous years, which also affected the recruitment process. Studies confirmed that there was a decline in rAOM/OME in children during the pandemic, possibly because of the preventative measures which decreased the incidence of infections (Aldè *et al.*, 2021; Allen *et al.*, 2022).

Descriptive analysis

A total of 42 children were assessed, 16 Males (mean age = 3.81 years, SD = 1.42 years) and 26 Females (mean age = 5.27 years, SD = 2.91 years). Seven children had unilateral OME (mean age = 6.14 years, SD = 2.73 years) and 35 had bilateral (mean age = 4.43, SD = 2.43 years). The total mean age was 4.71 (SD = 2.35) years, ranging from 2 to 11 years.

Regarding the scores of each question in the AOM6, the maximum score for all questions was 6 “very much a problem”, except for the speech question, where 9.5% of the parents (4 parents, all of which their children had bilateral OME) chose the score 7 “extreme problem”. Most scores of the six questions varied between a score of 3 and 4, except for the speech question where most parents scored 1 “not a problem”. Table 6.5 and Figure 6.1 display the scores of each question.

Table 6.5 Mean, SD, range, 95%CI of the scores and distribution of responses of each item in the AOM6 (Dark grey: most frequent response, Light grey: least frequent response.)

Questions	**Mean**	**SD**	**Min**	**Max**	**95% CI**		**Percentage of scores %**						
					Lower	**Upper**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Q1 Physical	3.33	1.34	1	6	2.93	3.76	14.3	9.5	26.2	31	16.7	2.4	0
Q2 Hearing	3.02	1.47	1	6	2.56	2.45	19	26.2	7.1	31	14.3	2.4	0
Q3 Speech	3.36	2.02	1	7	2.71	3.95	31	7.1	14.3	11.9	23.8	2.4	9.5
Q4 Emotional	3.43	1.36	1	6	3.02	3.86	7.1	23.8	14.3	33.3	19	2.4	0
Q5 Activity limitation	3.07	1.44	1	6	2.67	3.5	16.7	19	28.6	16.7	14.3	4.8	0
Q6 Caregiver Concern	4.45	1.55	1	6	3.98	4.93	7.1	14.3	31	28.6	4.8	14.3	0

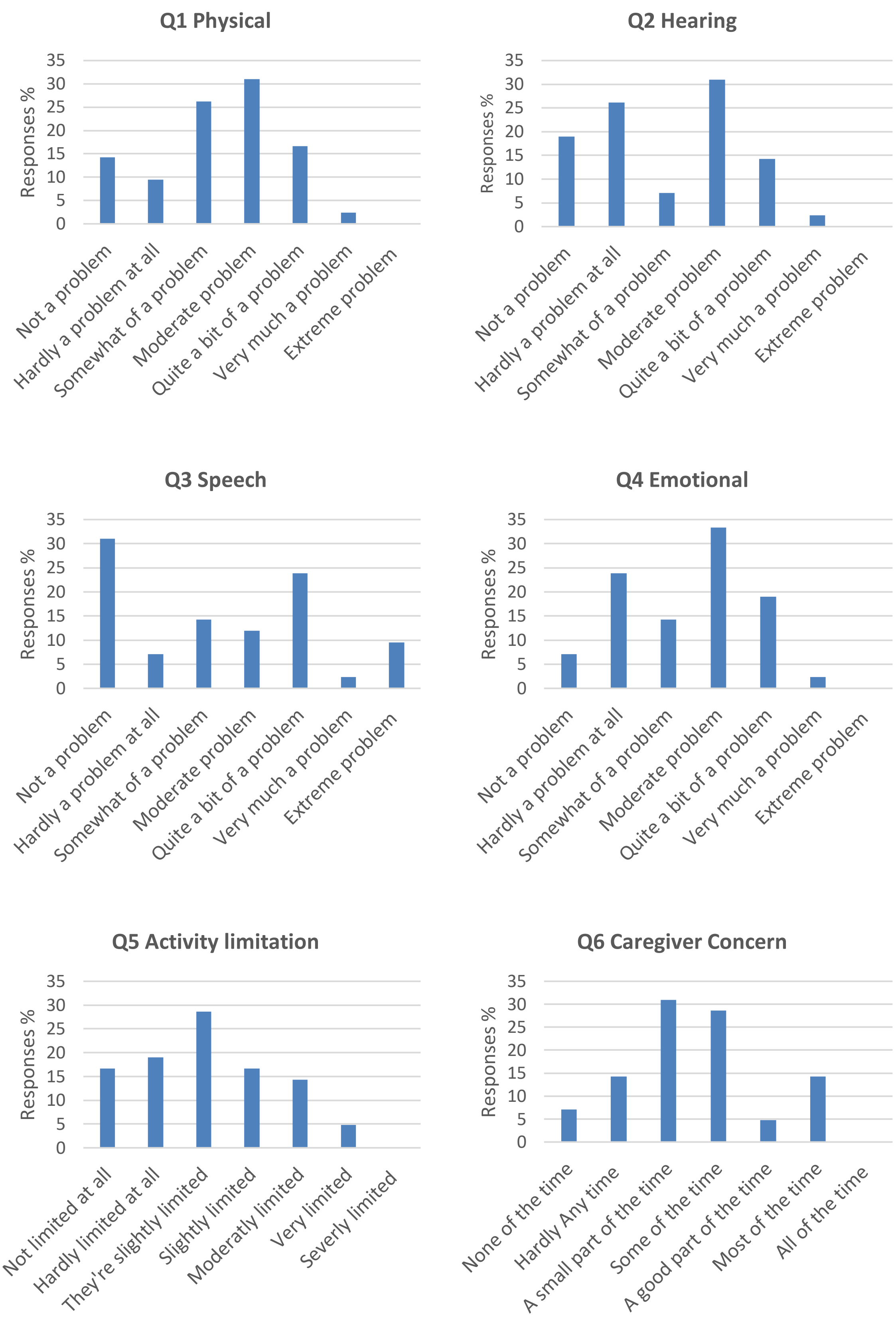


Figure 6.1 Bar charts of the percentage of each response in questions Q1 to Q6 of the AOM6

The AOM6 total for all n = 42 was 3.44 (SD =1.00) ranging between 1.5 and 5.33. The scores were lower in unilateral OME (n = 7, mean = 2.98, SD = 0.51), compared to those seen with bilateral OME (n = 35, mean = 3.54, SD = 1.05), but the difference between these scores was not significant ($p = 0.089$). The AOM6 HaS (mean = 3.19 SD = 1.61) showed better scores in unilateral OME compared to bilateral OME ($p < 0.001$). The AOM6 BaS had similar scores for both unilateral and bilateral OME (mean = 3.6) (Table 6.6).

Table 6.6 The scores of AOM Total, BaS and HaS in unilateral and bilateral OME.

Variable	n	M	SD	Mode	Min	Max	95CI%	
							Lower	Upper
AOM6 Total								
Total	42	3.44	1	4	1.5	5.33	3.14	3.73
Unilateral	7	2.98	0.51	3.17	2.17	3.67	2.5	3.45
Bilateral	35	3.54	1.05	4	1.5	5.33	3.18	3.9
Independent t-test results of the difference between Unilateral and Bilateral *p*-value = 0.089								
AOM6 HaS								
Total	42	3.19	1.61	1	1	6.5	2.69	3.69
Unilateral	7	1.71	0.49	1.5	1	2.5	1.26	2.17
Bilateral	35	3.49	1.59	1	1	6.5	2.93	4.03
Independent t-test results of the difference between Unilateral and Bilateral *p*-value < 0.001								
AOM6 BaS								
Total	42	3.57	1.02	3.75	1	5.25	3.25	3.89
Unilateral	7	3.61	0.85	4	2.5	5	2.82	2.4
Bilateral	35	3.56	1.06	3.75	1	5.25	3.2	3.9
Independent t-test results of the difference between Unilateral and Bilateral *p*-value = 0.54								

Regarding the effect of OM type on the scores, normality assumption was met for both infection type groups. Independent t-test showed that there was no significant difference between the AOM6 Total scores of children with rAOM and OME (Table 6.7).

Table 6.7 Independent t-test results of the difference between in AOM6 Total in rAOM and OME infection types

Variable	OME			AOM			*t*	*p*
	M	SD	n	M	SD	n		
AOM6 Total	3.46	1.07	31	3.39	0.77	11	0.19	0.424

Floor and ceiling effects

As mentioned earlier, the highest responses in all questions were a score of 6, except for Q3 (Speech), where 9.5% of the responses were the highest score (7). The lowest responses (1) were scored by 31% of the parents in Q3 (Speech), and in 16.7% in Q5 (Activity Limitation).

The lowest AOM6 Total score was 1.5 and the highest was 5.33, these ranges did not reach the minimum score of 1, or the maximum score of 7. A histogram of the scores of the AOM6 Total is displayed in Figure 6.2, showing a distribution that is close to normal, with Skewness = 0.19 and Kurtosis = -0.47.

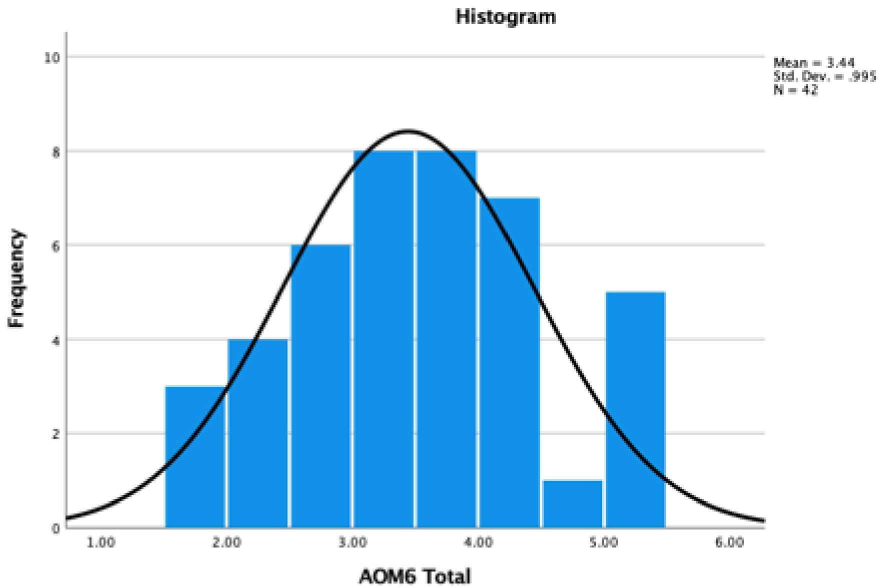


Figure 6.2 Histogram of the frequency distribution of the scores of AOM6 Total, with a normal distribution curve for reference

Internal Consistency

The items in the AOM6 Total, HaS and BaS had a Cronbach's α coefficient of 0.7, indicating good reliability. Table 6.8 displays the results of the reliability analysis.

Table 6.8 Coronach's α of the items in AOM6 Total, HaS and BaS

Scale	No. of Items	Cronbach's α	95%CI	
			Lower	Upper
All items	6	0.72	0.61	0.83
HaS	2	0.79	0.69	0.89
BaS	4	0.69	0.56	0.82

<u>Construct validity</u>

The results of the Pearson correlation between the scores of the AOM6 Total and VAS showed a strong negative correlation between the scores (r = -0.57, 95%CI (-0.75, -0.32), $p < 0.001$). These results suggested that the hypothesis constructed was correct, suggesting partially good construct validity.

6.5 Stage III: Assessment of the clarity and cultural acceptance of the AOM6

6.5.1 Aim

The aim of this stage was to ensure that the questionnaire was easy to read and follow, culturally acceptable and not offending.

This was achieved by analysing the results of the field test filled by parents during Stage II: Measurement of the psychometric properties of the AOM6 of the current experiment.

6.5.2 Method

The field test was composed of three questions that inquire about the AOM6 clarity and cultural acceptance. These questions are:

1- Do you agree that the instructions, questions, and answer choices of the AOM6 are clear?
2- Do you agree that the words used in the instructions, questions, and answer choices of the AOM6 are easy to understand and not ambiguous?
3- Do you agree that the instructions, questions, and answer choices of the AOM6 are culturally acceptable and not offending?

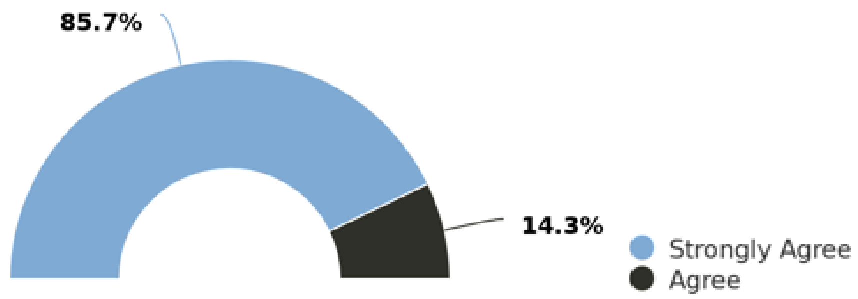
The response choices for this question were a in the form of a 5-point Likert scale: (1) Strongly agree, (2) Agree, (3) Neutral, (4) Disagree, and (5) Strongly disagree.

The frequency of the responses to these questions were calculated and displayed in graphs. In the literature, no specific percentages of responses were defined as a major problem, but it has been suggested that if more than 50% respond with ‘Disagree’ on an item, this item should be changed and translated back to English to confirm semantic equivalence (Hall *et al.*, 2018).

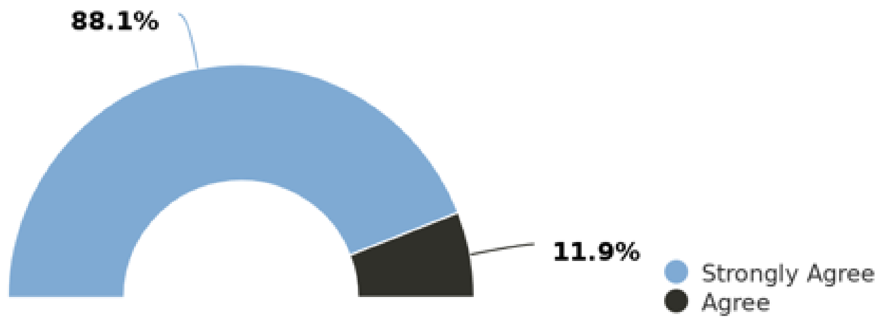
6.5.3 Results

The responses to the field test questionnaire enquiring about the parents' opinions about the AOM6 are displayed in Figure 6.3. Most parents strongly agreed that the AOM6 was a clear, easy to understand, and culturally acceptable questionnaire. None of the parents responded with neutral, disagree or strongly disagree.

(1) Do you agree that the words used in the instructions, questions and responses are clear?



(2) Do you agree that the words used in the instructions, questions and responses are easy to understand?



(3) Do you agree that the words used in the instructions, questions and responses are culturally acceptable and not offending?

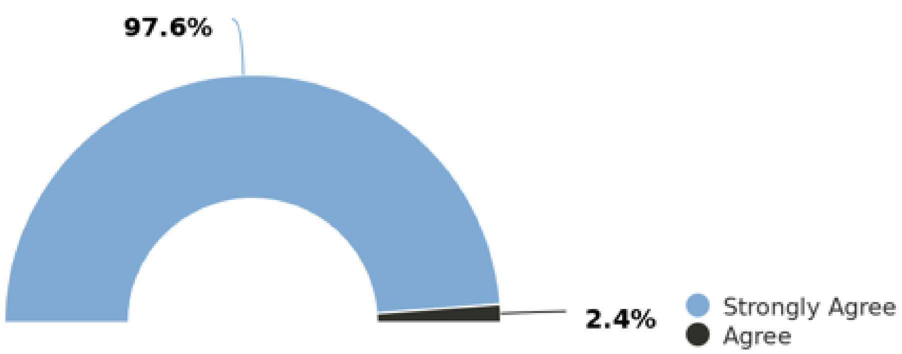


Figure 6.3 Charts of the response frequency for each question in the Field-Test of the AOM6

6.6 Discussion

The aim of this experiment was to develop a disease specific QoL questionnaire for OME in Arabic children. The OM6 questionnaire, which was developed by Rosenfeld et al. (1997), is a valid, reliable, and easy to fill disease-specific questionnaire (Saraf *et al.*, 2022) that has been translated and validated in different languages (Timmerman *et al.*, 2007; C.H. Heidemann *et al.*, 2013). The development of the Arabic version of the OM6 was done over three stages: Translation, measurement of the psychometric properties, and field testing.

Stage I: Translation and cross-cultural adaptation

In the first stage of the current study, the OM6 was translated from English to Arabic by following a series of procedures to achieve equivalence between the English and Arabic versions of the OM6. As mentioned earlier in the current chapter, assessing key equivalencies, including semantic item equivalencies, is important to ensure that the translation is not 'word for word' but rather a 'word for world' translation (Streiner *et al.*, 2015). The Arabic version of the OM6 was not very different from the English version, except for the speech item (item 3) where the words used in the question were replaced with slightly different phrases because the literal translation from English had a negative, exaggerated meaning. Proper cross-cultural adaptation can contribute to ensuring that the psychometric properties of a valid questionnaire are good across languages (Heidemann *et al.*, 2013). In a study that assessed QoL using the English version of the OM6 in non-English speaking population (Setswana – South Africa), 56 children aged 2-12 diagnosed with OME had a score of 1.67 (SD = 0.59) (Els and Olwoch, 2018). This score was not significantly different from that obtained from children without OME in the same study. This contradicted what previous studies have found where it has been shown that OM6 has good discriminative ability between the presence and absence of disease (Gan *et al.*, 2018; Tao *et al.*, 2018). The fact that the OM6 was administered in a non-native language and was translated on the spot may have affected the translation reliability and contributed to this finding.

Stage II: Measurement of the psychometric properties of the AOM6

The second stage of the current study was directed towards measuring psychometric properties of the AOM-6 questionnaire, namely the construct validity, internal consistency, and floor and ceiling effect.

The scores of each question as well as the mean score were calculated. The mean AOM6 Total for all participants was 3.44 (SD = 1, 95%CI: 3.14 – 3.73). Comparisons between the scores of current study and those obtained from previous OM6 validations studies, where the same inclusion and exclusion criteria of the study population were applied, are displayed in Table 6.9.

Table 6.9 Comparison between the scores of each item and total score between the current and previous studies

Item	Rosenfeld *et al.* (1997) English (US) n = 186	Kubba *et al.*, (2004) English (UK) n = 179	LAMEIRAS *et al.*, (2018) Portuguese n = 60	Current study Arabic n = 42
	Mean (SD)			
Q1 Physical	4.6 (1.8)	2.7 (1.7)	3.4 (2.1)	3.33 (1.34)
Q2 Hearing	2.7 (1.8)	3.6 (1.8)	3.8 (2.1)	3.02 (1.47)
Q3 Speech	2.3 (1.8)	2.2 (1.6)	2.7 (1.9)	3.36 (2.02)
Q4 Emotional	4.0 (1.8)	2.5 (1.6)	3.2 (2.0)	3.43 (1.36)
Q5 Activity limitation	3.4 (1.9)	2.3 (1.7)	2.6 (1.9)	3.07 (1.44)
Q6 Caregiver Concern	4.9 (1.7)	3.1 (1.8)	4.1 (2.4)	4.45 (1.55)
Total mean score	2.8	2.7 (1.7)	3.3 (1.47)	3.44 (1)

The Portuguese and Arabic version total scores were very similar, and both were higher than the scores of the English version. This finding should be assessed with caution because both the Arabic and Portuguese studies had much smaller sample sizes (n < 70) than the validation studies on the English version (n >100), but different scores could be accepted due to the cultural differences (Kubba *et al.*, 2004) and possible differences in severity of the conditions.

In the current study, the most common response in almost all questions (except for the speech item) varied between “somewhat of a problem” (or what is equivalent to a score of 3) to “moderate problem” (equivalent to score of 4). The most common response to the speech item (item 3) was a score of 1 (not a problem/ not present) with a frequency of 31%. On the other hand, this same question was the only item where parents (9.5%) gave a score of 7 (extreme problem). Similar findings were observed in the study by (Rosenfeld *et al.,* 1997), where the most common response (57%) to the speech item was a score of 1. This observation has been addressed by Heidemann *et al.* (2013), where they suggested that parents found it difficult to evaluate speech impairment because most children in their study were very young and still in their early stages of language development. The results from the current study also suggested that speech evaluation is a subjective observation that depends on the parents’ knowledge of what constitutes normal speech development, which could have led parents to respond mostly at the extreme ends of the Likert scale. A recent study in Saudi Arabia showed that parents lacked the ability to detect speech delay in their children (Alakeely *et al.*, 2022), which could explain the variations and wide range in responses in the speech item of the AOM6.

The highest score in the AOM6 was seen with the caregiver concern question (mean= 4.45, SD = 1.55). Previous studies also found that the caregiver concerns scored the highest among other questions (Rosenfeld *et al.*, 1997; Lameiras *et al.*, 2017), which could be expected with most parents where the health and wellbeing of their children is a priority, and the illness of their child could be a source of anxiety and concern.

Bilateral OM had poorer AOM6 total scores (mean = 3.54, SD = 1.05) compared to unilateral OM (mean = 2.98, SD = 0.51), but this difference was not significant (p = 0.089), suggesting that even unilateral OME could be severe enough to affect OME just as bilateral OME could. In their study on how appropriate the OM6 is as a discriminative questionnaire, Kubba *et al*. (2004) found that bilateral OME yielded the worst scores compared to other conditions, including cases where tympanogram types were other than type C2 or B. Regarding the effect of infection type, there was no significant difference in AOM6 Total scores between rAOM and OME (p = 0.424) which could be expected since the two types tend to overlap, because rAOM could sometimes present with effusion (Heidemann *et al.*, 2013). The Danish version OM6 showed an ability to discriminate between infection types (Heidemann *et al.*, 2013). In their study, participants were divided into children with OME and rAOM and those with OME and no rAOM, and the OM6 showed a significant difference in the scores between the two groups, which could be explained by the fact that rAOM could lead to more severe symptoms (including ear ache and emotional distress) (Heidemann *et al.*, 2013). In the current study the comment was only on the difference in AOM6 Total between rAOM and OME. Therefore, the possibility that a good portion of children had OME associated with rAOM, or that parents were not certain about the frequency of infections in the past, rendered the scores similar in both groups. The study by Tao *et al.* (2018) produced receiver operating characteristic (ROC) curves to assess the discriminative validity of the OM6 and its subgroups for rAOM, COM, as well as different levels of severity of OM defined by history of episodes of AOM and duration of OME episodes. The area under the curve (AUC) of the ROC for COM and rAOM was approximately 0.6, which was considered a less than optimum ability to discriminate (0.5 indicated no ability to discriminate). The AUC for the different levels of severity was even less than that for the OM subgroups. Similar to the study by Tao *et al.* (2018), the current experiment is a cross sectional study, which posed an issue on the validity of the diagnosis because of the high dependance on the parents' history of OME, because the information about the number of episodes and the duration of OME was provided by parents based on what they have been told in other centers. Nevertheless, it has been suggested that the OM6 is able to discriminate between presence and absence of OME, but it was not considered an optimal measure of clinical severity (Timmerman *et al.*, 2007; Tao *et al.*, 2018) .

Although the knowledge of the scores alone in the current study would not provide information about whether the AOM6 can assess QoL, because there is no control group of normal children or post-intervention group, the similarities with previous studies suggest that AOM6 could be considered for assessment of QoL in Arabic children with OME. The scores of the AOM6 provide a baseline of how OME affects QoL and were analysed to assess the psychometric properties.

There were no floor and ceiling effects in the AOM6 Total. Previous studies of the OM6 also confirmed the absence of this effect (Brouwer *et al.*, 2007; Heidemann *et al.*, 2013). Larger sample sizes with different severity degrees of OM might show some floor effect in less severe OM (Heidemann *et al.*, 2013).

Construct validity was partially assessed by testing the theoretically derived hypothesis based on previous knowledge that the global QoL of the child would be strongly and negatively affected by the disease specific FHS of the AOM6 (Heidemann *et al.*, 2013), which in the current study was confirmed by the strong negative correlation between VAS scores and AOM6 Total (r = -0.57, 95%CI (-.75, -.32), p <.001). Previous studies assessed the construct validity by measuring the correlation between the OM6 and other questionnaires such as the VAS (Rosenfeld *et al.*, 1997), Caregiver impact questionnaire (CIQ) (Heidemann *et al.*, 2013) and Pediatric QoL (PedsQL) (Tao, Schulz, Donna B. Jeffe, *et al.*, 2018). Similar findings to the current study were found in the study by Rosenfeld *et al*. (1997) (r = -0.64, p < .001). The Danish version (Heidemann *et al.*, 2013) showed a small negative correlation between the two measures (r = -0.33, p < .001).

The internal consistency is part of the measurement of the reliability of a QoL questionnaire (Gan *et al.*, 2018), and is assessed by Cronbach's α. The Cronbach's α coefficient for all factors (AOM6 Total, BaS and HaS) was approximately 0.7, which was considered an indicator for good reliability. Previous studies also confirmed good reliability of the OM6, not only in the English version, but also in other languages. An example of these studies was a large study on the Danish OM6 by Heidemann *et al.* (2013), where the Cronbach's α was 0.85, 0.9 and 0.76 for OM6 Total, BaS, and HaS components, respectively. Additional analysis in the study by Heidemann *et al.* (2013) was performed using PCA, where they found that covariation between the hearing and speech items, produced a perfect fit to the model, which suggested closeness of these items to each other. One of the reasons behind the decision of the authors in the aforementioned study to set these two items (hearing and speech) in the model as covariates was because of the variations in responses of these two items. In the current study, the factor analysis was not performed due to the smaller than required sample size, therefore the choice of components was based on previous studies

(Heidemann *et al.*, 2013; Tao *et al.*, 2018). The results from the current study thus suggested that the AOM6 has partially good reliability, but to ensure all aspects of reliability, test-retest reliability should be measured. Test-retest reliability was assessed as part of Experiment 4 (Section 7.5.3).

Stage III: Assessment of the clarity and cultural acceptance of the AOM6

Stage III aimed to ensure the clarity and cultural acceptability of the questionnaire. Most parents strongly agreed that the AOM6 was clear, easy to understand and not offending.

6.7 Conclusion

The main aim of the study was to develop the AOM6, which is the Arabic version of the OM6 questionnaire. Developing a QoL questionnaire by translating an existing questionnaire in a different language requires cultural adaptation, assessment of its clarity and cultural acceptance by the target population, and assessment of its psychometric properties.

<u>Main conclusions of this study were:</u>

- Cross-cultural adaptation is an important step when translating a QoL questionnaire because differences in languages can affect equivalence. The words in the speech item in AOM6 were slightly changed to achieve item and semantic equivalence.
- The AOM6 total scores were similar to those obtained from previous studies.
- The AOM6 total scale and subscales (HaS and BaS) had good internal consistency, indicating good reliability.
- The AOM6 had no floor and ceiling effect, and potentially good construct validity.
- There was a wide range of responses in the speech item, indicating that parents may not be able to correctly assess their children's speech development. Therefore, it is suggested that parents should be educated about the normal speech and language developmental milestones.
- Most parents strongly agreed that the AOM6 is a clear, easy to understand, and culturally acceptable questionnaire.

Test-retest reliability and other aspects of construct validity, such as the relationship between the AOM6 and SRTs, were measured in Experiment 4 (Chapter 7). Other psychometric properties, such as responsiveness, did not take place during the current PhD because they require a setting of an RCT or a prospective observational study of an intervention, which was not part of the study design of the main PhD aim, but can be tested in the future. Future studies will also be directed towards assessing discriminant validity, where children with different middle ear statuses can be compared.

Chapter 7 Experiment 4: Measuring speech recognition and QoL using the PAAST SiQ and Arabic OM6, respectively, in Arabic children with OME

7.1 Introduction

Otitis Media with effusion is the most common cause for CHL in children (Klein and Pelton, 2018) and could affect language development if not managed adequately (Cai and McPherson, 2017). This condition is also associated with poor QoL (Brouwer *et al.*, 2005).

The outcome measures utilised in researching the effects of OME on Arabic-speaking children in Saudi Arabia are clinical ear examination using otoscopy and tympanometry, as well as conventional hearing assessments using narrowband stimuli such as Pure Tone Audiometry (PTA) (Al-humaid *et al*., 2014). One limitation with PTA assessments is their unsuitability for many 3 year olds, despite this age group being more susceptible to OME than older children (Martines *et al.*, 2011). Another shortcoming of the methods currently used in Saudi Arabia is their inability to assess the ability of a child to recognise speech, which would indirectly reflect their speech and language development, as well as their QoL. Consequently, it is not clear how OME affects speech intelligibility and QoL in Arabic-speaking children. According to a review by Homøe *et al*. (2020), it is strongly recommended to measure both hearing and QoL in order to assess the outcome of OME management.

Speech recognition tests can be used to assess hearing in children aged 2.5 years and older, especially if the test is easy, engaging, automated and readily available on smart devices, such as the Automated McCormick toy test (AMTT) (Hall *et al*., 2007). An Arabic version of the MTT, the Paediatric Arabic Auditory Speech Test (PAAST), was developed by a former PhD student in the University of Southampton, to assess hearing in noise (SiN) in Arabic-speaking children with SNHL (Al-Kahtani, 2020). In the current study, the PAAST was applied to explore speech recognition in quiet (SiQ) in Arabic-speaking children with OME. Currently, there are no data on SiQ in Arabic children aged 3.5-6 years with confirmed OME, therefore, the PAAST SiQ was assessed for its ability to detect OME-related CHL in this group of children. Previous studies in the current PhD ensured homogeneity of the speech material of the PAAST SiQ, and its sensitivity to OME-related SCHL in NH adults (Chapter 5).

Health-related quality of life (HR-QoL) is another important aspect to consider. There are currently no Arabic tools for assessing the QoL in children with OME.

An Arabic QoL tool was developed in the current PhD by translating and culturally adapting the OM6 from English to Arabic and assessing aspects of its validity (Chapter 6). The AOM6 was culturally acceptable and easy to understand by more than 97% of parents. It exhibited good internal consistency, partially good construct validity and negligible floor and ceiling effect. In the current study the Arabic OM6 was assessed for its test-retest reliability and its correlation with speech recognition in the Arabic-speaking children with OME.

Children with OME can be categorised by type, either unilateral or bilateral OME. The ears of children with OME can be categorised based on status (OME and No OME), and laterality (whether the ear belonged to a child with unilateral or bilateral OME). Children with unilateral OME have one OME ear on one side and No OME ear on the other, whereas in bilateral OME, both ears are of OME statuses. OME ears tend to have high PTA hearing level thresholds (PTA-HL) and increased SRTs, compared to No OME ears (Brown *et al.*, 2019). OME type can affect the levels of SRT, which tend to increase in bilateral OME as opposed to unilateral if measured binaurally with speakers (Hall *et al*., 2007). Effect of the OME type on QoL in this group of children varied between studies and type of questionnaire used, but in general, bilateral OME leads to increased (worse) scores of OM6 compared to unilateral OME. The relationship between hearing and QoL was assessed in several studies, but the study designs and tools used varied between these studies, leading to variable results (Rosenfeld *et al.*, 1997; Heidemann *et al.*, 2013). None of these studies assessed the relationship between speech recognition and QoL in OME children.

This chapter is directed towards addressing Aim 4 in the current PhD project, by exploring the effect of OME on speech recognition and QoL in Arabic children, to assess whether the PAAST SiQ and AOM6 could be considered valid, reliable outcome measure tools in this group of children.

7.2 Research question, aims, and objectives

To say that a particular tool can measure the intended outcome, it is essential to examine all aspects of the validity of that tool, which often requires extensive research. The scope of this study allowed to investigate the effect of OME on these tools and measure aspects of the discriminative validity and the reliability of the PAAST SIQ and AOM6.

The following research question was formulated:

Can speech recognition and QoL using the PAAST SiQ and AOM6, respectively, be considered as valid, reliable outcome measures in Arabic children with OME?

The aims directed towards answering the question are:

Aim 1: To investigate the effect of OME on speech recognition and QoL

Aim 2: To investigate the relationship between speech recognition and QoL

Aim 3: To assess the reliability of the of the PAAST SiQ and AOM6

The objectives directed towards achieving these aims are:

Objectives directed to achieve Aim 1

1. To measure the effect of OME status, laterality, and age on SRTs.
2. To measure the ability of SRTs to predict PTA-HL.
3. To measure the effect of OME on the AOM6.

Objective directed to achieve Aim 2

To measure the relationship between the AOM6 and SRTs and PTA-HL.

Objectives directed to achieve Aim 3

1. To assess the test-retest reliability of the PAAST SiQ.
2. To assess the test-retest reliability of the AOM6.

7.3 Method

7.3.1 Participants

The method of sampling used was convenience sampling. Parents of participants who fit the criteria were approached in the ENT clinic of King Abdulaziz University Hospital (KAAUH). Sample size calculation was based on the null hypothesis (H0) that there is no correlation between SRT and PTA-HL or no effect of OME status on SRT, with an α-error probability of 0.05 and 1-β error probability of 0.8. The number of participants needed to measure the correlation between the speech recognition and QoL was calculated based on H0 that there is no correlation between speech recognition and QoL, with a standard deviation (SD) of 8.9 dB of SRTs in children and SD of 1.5 for the parents' responses to the hearing item in the AOM6. G*power calculator was used to obtain the sample size (Faul et al., 2009), and the resulting sample size was 64 participants. Given the Covid-19 constraints, n= 50 participants were recruited. The inclusion and exclusion criteria are mentioned in Table 7.1.

Table 7.1 Inclusion and exclusion criteria for participants

Inclusion	• Children aged 3-6 years and their parents
	• Native Arabic Speakers
	• Chronic Otitis Media with Effusion (OME)
	    • Presence of effusion in the middle ear 3 ≥ months in
	    • Tympanometry: Type B or C2
	    • Unilateral or bilateral

Exclusion	• Surgical treatment for the otitis media in the past 3 months • Syndromes/oral facial anomalies e.g., cleft palate • Other concurrent diseases that can affect the QoL e.g., heart or lung disease • Children diagnosed with permanent SNHL or CHL

Although the initial study design focused on recruiting children aged 3-5 years, the age group was extended to 6 years as a maximum age, due to the paucity of children attending hospitals during the COVID-19 restrictions, and the possible decrease of OME prevalence in children during the pandemic (Aldè *et al.*, 2021; Allen *et al.*, 2022), which had an effect on recruitment. This change in study design allowed for collecting data from the 50 children within the set timeframe. The mean age and SD of the participants are included in the descriptive results. Also, the separation of the sample into unilateral and bilateral was not planned.

7.3.2 Procedure

The study took place between December 2021 and April 2022. Participants were recruited from the ENT clinic where they have been examined by an ENT specialist to confirm the presence of OME in at least one ear by clinical examination using otoscopy as well as tympanometry, where Type C2 or B confirmed the presence of fluids behind the ear drum (Rosenfeld and Kay, 2003). The centres where the test took place were King Abdulaziz University as well as a private clinic (Irfan clinic) in Jeddah, Saudi Arabia. The data collection was performed entirely by the researcher.

The procedure was composed of two main parts: Part 1 - measuring hearing and speech recognition thresholds of children using PTA and PAAST SiQ, respectively, and Part 2 – measuring QoL by obtaining responses from parents on the Arabic OM6.

7.3.2.1 Part 1: PTA and PAAST in children

This part took place in a soundproof room in the aforementioned centres. The background noise in the room was monitored using the SLM. The maximum background noise did not exceed 40 dBA, which is within the maximum acceptable ambient noise level in the testing room when screening for air-conduction thresholds (Gelfand, 2009). The child was seated in a comfortable age-appropriate chair during these sessions. This part took take 30-35 minutes with breaks.

PTA

Hearing thresholds of the child were tested using the standard clinical procedure for PTA in accordance with the British Society of Audiology recommended procedures (BSA, 2018). The child wore TDH49 supra-aural headphones, heard pure tones of different frequencies and intensities and was instructed to raise their hand if they heard them, even if the tone was very low. The PTA was performed at frequency ranges from 0.25-8 kHz monaurally (each ear) with a maximum starting level of 60 dB HL. The air-conduction (AC) thresholds were obtained at frequencies 0.5, 1, 2, and 4 kHz in the right ear and then the left ear, the 0.25 frequency was tested afterward in the right ear, followed by the left ear. This ensured that the PTA thresholds at frequencies 0.5, 1, 2, and 4 kHz were obtained, as the average AC thresholds at these frequencies can provide acceptable information about the hearing of a child with OME (Chow *et al.*, 2019) in case the child was unable to complete the test at the remaining frequencies. The preparation of this test and the procedure lasted 15 minutes.

PAAST SiQ

After the child had completed the PTA testing, they were tested with the PAAST SiQ to measure their speech recognition threshold in quiet (SRT SiQ) in each ear (monaurally). The PAAST SiQ was designed as a code in MATLAB which was originally designed by Dr. Daniel Rowan and modified by Dr. Ben Lineton. The test consisted of a graphical user interface (GUI) containing 10 pictures (Figure 7.1), and headphones (Sennheiser HD 650) connected to a Babyface Pro sound mixer and the computer and placed over the child's ears.

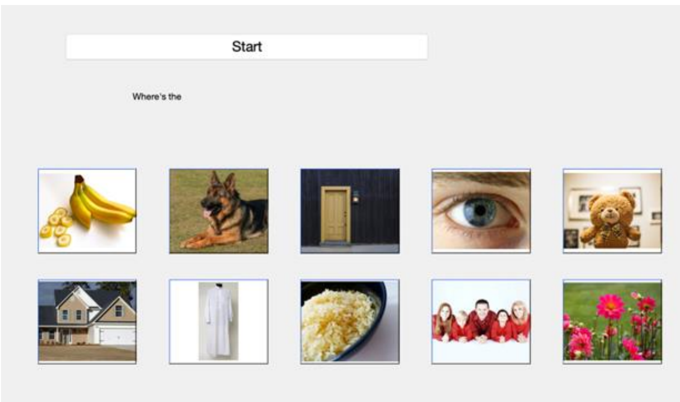


Figure 7.1 The graphical interface of the PAAST SiQ - MATLAB version

The tester first ensured the child's familiarity with all words by naming each picture and asking the child to point to the named picture. Afterwards, the child was informed that they would listen to sentences in the form of ("where is the (WORD)?") through the headphones and were asked to touch the picture that corresponded to the heard word, and the researcher chose that picture using the mouse connected to the computer. Children were asked to guess if they were not sure what the presented word was.

The method used to obtain the speech recognition threshold (SRT) was a 2-down/1-up adaptive procedure, first described by Levitt (1971). This method estimates the monaural threshold corresponding to 70.7% correct responses by increasing the stimulus level following each incorrect response but not decreasing it until two correct sequential responses occur (Kingdom and Prins, 2010). The initial level (IL) was set to 50 dBA. This IL value was based on a previous experiment by the investigator, where the PAAST SiQ was applied in an adaptive procedure on adults who were tested under OME-related SCHL conditions (Section 3.2.4.2). The measure resulting from the test is the speech recognition threshold (SRT) at 70.7% correct (Figure 7.2).

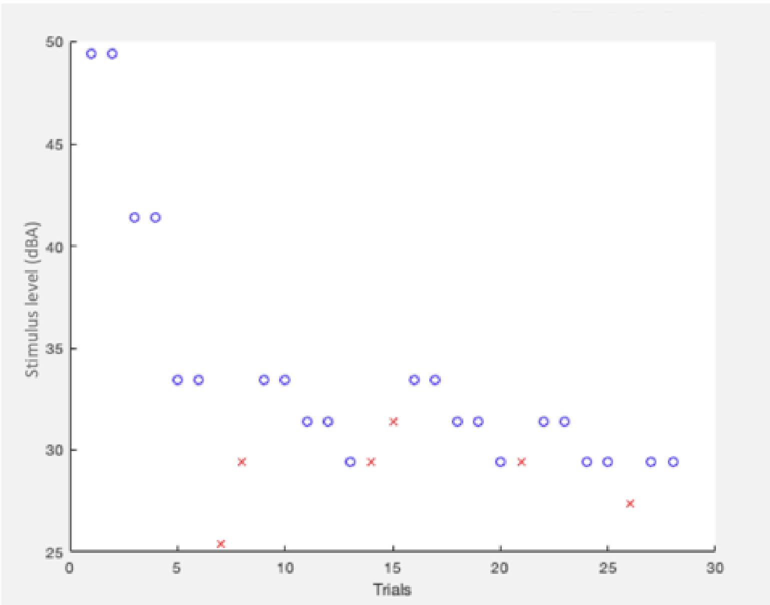


Figure 7.2 An example of a staircase track of the 2-down 1-up adaptive procedure of one the participants

The duration of one complete adaptive procedure using the PAAST SiQ was approximately 3 minutes. The test was administered three times, the first time on the OME ear if the participant had unilateral OME, or the right ear in bilateral OME, the second time on the other ear, and the third time a repeat of the ear with the better SRT. The reason why the better ear was chosen to assess repeatability was to ensure the best results, and because previous studies used the better ear for assessing repeatability (Al-Kahtani, 2020). The ear with OME was tested first to ensure getting results on the effect of OME on SRT, in case the child was unable to complete all tests. The child was given a break between tests. Total testing time with the PAAST SiQ took up to 15-20 minutes including preparation, familiarisation, and breaks.

7.3.2.2 Part 2: Arabic OM6 from parents

The participants' parents were asked to fill an electronic form of the AOM6 (Appendix G.3) on the researcher's iPad on the same day of the test (Repeat 1). They were also requested to fill an electronic form of the AOM6 questionnaire 7 days later (Repeat 2), the form was sent to parents via a WhatsApp message. Each child was assigned an ID number by the researcher, which was provided to the parents. The parents were asked to enter the ID number assigned to their child at both times to match the first and second repeat. The rationale behind test repetition after this time period was to measure the reliability of the test while ensuring not much change occurred to the status of the ear (Rosenfeld et al., 1997).

7.3.3 Materials and calibration

Similar to section 3.2.2.6.

7.3.4 Ethics and Safety

Ethical approval (ERGOII 68934) for this experiment was granted from the University of Southampton on 9 December 2021 (Appendix I.5). Ethical approval for the application (21-578) to King Abdulaziz University was also approved on 6 December 2021 (Appendix I.8). The speech never exceeded 60 dB A and was primarily presented near absolute threshold. This is well below the maximum safe level for 8 hours of exposure per ear per day as outlined in The ISVR Report 808.

This experiment was conducted during the COVID-19 Pandemic, shortly after lifting lockdown mandates. The investigator followed Saudi Ministry of Health (MOH) guidance on prevention of contracting COVID-19 (MOH, 2020). Precautions including social distancing, regular handwashing, and face masks for both the investigator and participants were taken. All equipment, tables, and chairs where sanitised using an antibacterial swab/spray between each participant.

7.4 Analysis Strategy

Statistical analysis was performed using SPSS Version 28.0. A statistician was consulted about Aim 1 analysis.

7.4.1 Aim 1 To investigate the effect of OME on the speech recognition and QoL

Objective 1 To measure the effect of OME status, laterality, and age on SRTs

The descriptive analysis was directed towards presenting the average SRTs and PTA hearing thresholds levels (PTA-HL) across the right and left ears. The PTA-HL are the average threshold levels across frequencies 0.25 – 4 kHz for each ear.

The current study tested participants using the PAAST monoaurally, providing SRTs for the right and left ear. Participants varied in their OME conditions, these conditions were OME status of the ear (whether the ear had OME and No OME) and laterality of the OME (whether the OME ear belonged to a unilateral or bilateral OME), bearing in mind that each participant had one OME laterality status, and two OME statuses, one for each ear. The fact that each participant had two SRT measurements (except for those who completed one PAAST repeat in one ear) rendered the data from ears not independent from each other.

Given this information, to analyse the effect of OME conditions and age on speech recognition, a model that took into account the multiple levels of data (multilevel model) was the most suitable method, namely Linear mixed model (LMM). Linear mixed models are an extension of simple linear models that allow both fixed and random effects. These models are used in studies where there is non-independence in the data (Twisk, 2019), such as the data in the current study where the right and left ear have different measurements for the same participant. Random effects represent the regression coefficients that are being modelled from the sample, which also refers to the randomness in the probability model for the group-level coefficients (Twisk, 2019). In the case of the current study, participants were set as the random effect.

Fixed effects corresponded to parameters that do not vary, or in other words, effects that lead to regression coefficients that do not vary by group (Gelman and Hill, 2007). In the current study the OME laterality and status were the fixed effects, with age considered as a covariate. This model aimed to measure the:

1. Difference between a No OME ear and an OME ear in a participant with unilateral OME.
2. Difference between an OME ear in a subject with unilateral OME and an OME ear in a subject with bilateral OME.
3. Average change in SRT with age.

Objective 2 To measure the ability of SRTs to predict PTA thresholds

Linear Mixed Model was conducted to assess the relationship between SRT and PTA-HL, where the PTA-HL was the dependent variable. The fixed factors were OME status and OME laterality, both controlled for SRTs. this analysis allowed to achieve the following objectives:

1- To assess the relationship between PTA-HL and SRTs, taking into consideration the OME conditions of the participants
2- To assess the effect of age on the relationship between PTA-HL and SRTs
3- To obtain the formula that would allow for prediction of PTA-HL from SRTs

Also, analysis was performed to assess the relationship between SRTs and PTA-HL average of the following frequency combinations:

1- PTA average 0.5 and 1 kHz (dB HL)
2- PTA average 1 and 4 kHz (dB HL)
3- PTA average 0.5, 1, and 4 kHz (dB HL)
4- PTA average 0.5, 1, 2 and 4 kHz (dB HL)
5- PTA average 0.25, 0.5, 1, 2 and 4 kHz (dB HL)

Different combinations were assessed to explore whether lower frequency combinations have different relationship with SRTs compared to broader frequency combinations, given that previous studies showed that the SRT SiQ correlates strongly with PTA frequency combination 0.5-1 kHz in adults (Smoorenburg, 1992). The frequency combinations chosen to be examined in the current study were based on combinations assessed in a similar fashion, using the AMTT (Palmer *et al*., 1991; Summerfield *et al.*, 1994; Hall *et al.*, 2007). This step in the analysis allowed for comparison between the current study and previous studies that investigated the relationship between SRTs and PTA-HL average of different frequency combinations. The frequency combination 0.25, 0.5, 1, 2, and 4 kHz in the current study is an additional combination that was not examined in the studies mentioned above and was added to assess the effect of low-frequency threshold (0.25 kHz) on the relationship with SRTs, given that OME is usually associated with rising, low-frequency CHL (Cai *et al.*, 2018; Chow *et al.*, 2019).

The relationship between the average PTA-HL (right and left ear) of each of the five frequency combinations and average SRT (right and left ear) was assessed by conducting Pearson correlation analysis.

<u>Objective 3 To measure the effect of OME on the AOM6</u>

This section described the results of the total scores of the AOM6 (Total) and the AOM6 (HaS), including the scores categorised by laterality (unilateral and bilateral). The method of applying the questionnaire and scoring the results were similar to those used in Sections 6.4.2 and 6.4.3. An independent t-test looked at the difference in the scores between unilateral and bilateral OME. This analysis was done based on the prediction that bilateral OME would affect QoL more than unilateral OME would.

7.4.2 Aim 2 To investigate the relationship between speech recognition and QoL

<u>Objective: To measure the relationship between the AOM6 and SRTs and PTA-HL</u>

As mentioned before, AOM6 measures QoL in children with OME through two components: behavioural and symptomatic aspects (BaS), which are assessed through questions on pain, behavioural changes, activity limitations, and caregiver concern, as well as hearing and speech aspects (HaS) answered through the two corresponding questions.

The relationship between hearing and QoL was investigated by assessing the following relationships:

1- SRT and AOM6 (Total)
2- SRT and AOM6 (HaS)
3- PTA-HL and AOM6 (Total)
4- PTA-HL and AOM6 (HaS)

The Pearson correlation analysis was conducted for these relationships using the average SRTs and PTA-HL (Average right and left ears, n= 50), and first repeats of AOM6 (n = 50) to ensure that the QoL represented the status of participant at the time of the hearing test.

To explore the relationship between parents' perceptions of their children's hearing abilities and measured speech recognition, a Spearman correlation was applied to test the relationship between SRTs and the 7 Likert scale responses from the Hearing question in the AOM6 (AOM6 (HI)). Similar analysis was performed for the PTA-HL average. This type of correlation was conducted because the analysis was performed between continuous (average SRT and PTA-HL, n=50) and ordinal (AOM6 HI, n= 50) variables.

7.4.3 Aim 3 To assess the reliability of the of the PAAST SiQ and AOM6

Objective 1 To assess the test-retest reliability of the PAAST SiQ

Test-retest reliability was assessed in participants who repeated the test a second time for the better ear (n=32). This sample included five children from the pilot study to allow the opportunity to measure reliability on a larger sample, two children were within the age range of the study, and three ranged between 6.5 and 7 years. After participants performed the PAAST SiQ in the right and left ears, they were immediately (a maximum of 5 minutes break) tested again in the ear with the better SRT. The test-retest reliability was assessed between the ear with the better SRT (Repeat 1) and the repeat of the same ear (Repeat 2).

When measuring reliability, systematic variations and random errors must be considered. Systematic variations can occur due to certain factors that influence all subjects' measurements such as learning effect or fatigue, whereas random error represent changes in within-subject measurement that are not explained by certain factors (not systematic) (Bell, 2001). A repeated measure ANOVA was carried out to measure the mean difference in SRTs (dependent variable) between repeats (independent variable with two levels), providing information about the systematic variations. Variability was measured by within-subjects standard deviation (SDω) of the scores (Summerfield *et al*., 1994), as well as repeatability which can also be calculated from the SDω. Repeatability defines the minimum significant difference between scores when the test is repeated under identical conditions, and it could be useful to assess the change in SRTs with an intervention or hearing aids (Lovett *et al*., 2013).

Assuming the data was normally distributed, and a single score was within ±1.96 SDω of the mean with a probability $P \geq 0.95$, the random error could be assessed by calculating the variability. The smaller the SDω the better the reliability of the test (Summerfield *et al.*, 1994; Bland and Altman, 1996). Another measure of random error is correlation coefficient. The intra-class correlation coefficient (ICC) is the proportion of the variability in the measured quantity that is due to the variability in the quantity between subjects. If the individual-to-individual variability was a large proportion of the total variability, the test can be regarded as reliable. The ICC is a value between 0 and 1: values greater than 0.75 constitute good reliability (Koo and Li, 2016).

Objective 2 To assess the test-retest reliability of the AOM6

Analysis similar to the PAAST SiQ was performed for the AOM6, where n=28 participants were assessed for the test-retest reliability. The assessed score was the AOM6 (Total) which was performed on the day of the test (R1) and then 7 days later (R2).

7.5 Results

A total of n = 61 potential participants were approached for this study, six of which were a pilot of NH children whose data were included for correlation with PTA hearing levels (HL) and test-retest reliability of the PAAST SiQ. Five of the approached participants who fit the inclusion criteria were not included in the study, two of which did not complete neither PTA or PAAST, and three children were not attentive or too shy to complete the PAAST but were able to perform PTA. Participants who were included in the data were 50 children (Females: n= 19, Males: n= 31), with mean age of 5.0 ± 0.7 years. All participants and their parents were native Arabic speakers, but of different nationalities including 40 Saudis, 7 Egyptians, 2 Syrians, and 1 Lebanese. Participants who completed all three PAAST SiQ repeats were n= 27, participants who completed two repeats (right and left ears) were n = 20, and n = 3 (ages 4.6, 5.9 and 4.3 years) only completed one repeat because they were tired. A summary of the participants is included in Figure 7.3.

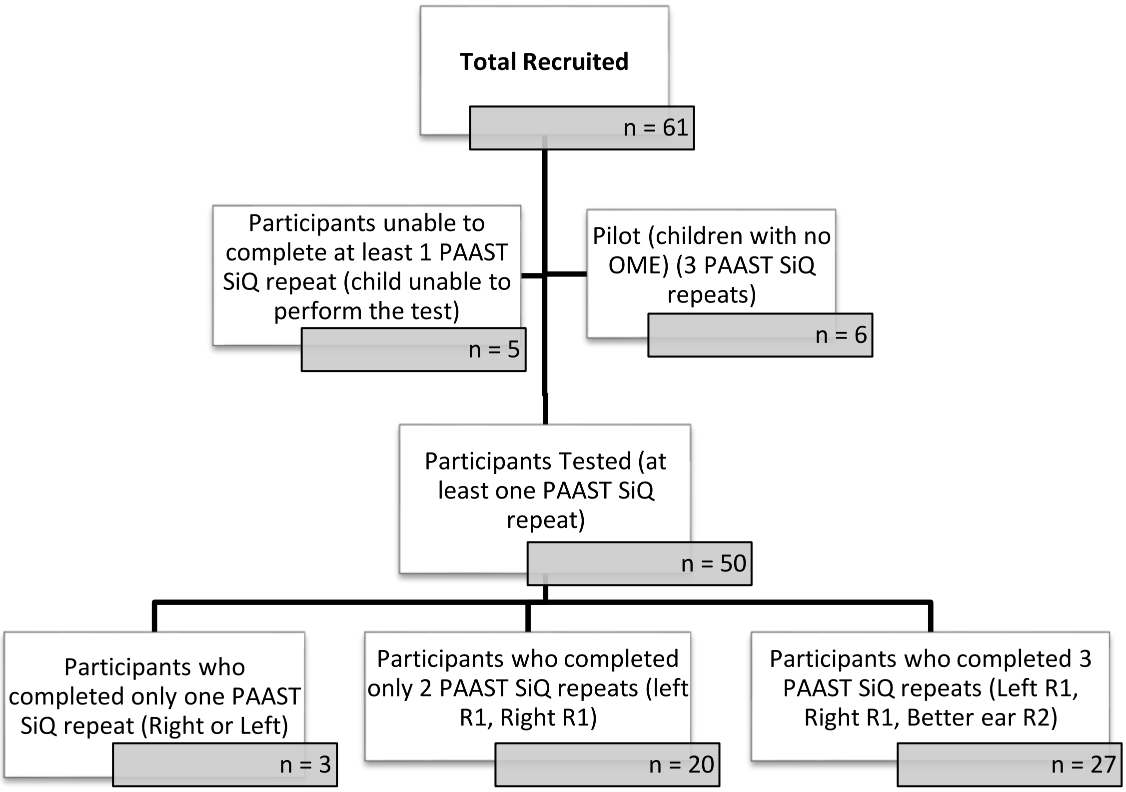


Figure 7.3 Flowchart of the participants included (and excluded) in the study.

All participants' parents completed the first repeat of AOM6, but only n=28 completed the AOM6 the second time. Of the n=31 children whose parents did not fill the AOM6 a second time, n=15 were due for an ear grommet insertion surgery within the seven days period, therefore, they were not asked to fill the questionnaire again because the possibility of a change in responses would be higher. The remaining parents did not respond to the request to fill out the form a second time.

7.5.1 Aim 1 To investigate the effect of OME on speech recognition and QoL

Objective 1 To measure the effect of OME status, laterality, and age on SRTs

Descriptive analysis

Descriptive analysis was conducted to present the data of SRTs and AOM6 scores in children with OME. Table 7.2 presents the number of children with unilateral and bilateral OME, and the mean age in each group.

Table 7.2 Number of children with unilateral and bilateral OME, percent of total, and their mean age and SD

OME Laterality	n (participants)		% of Total	Mean Age (years)	SD (years)	Age range (years)
Unilateral	**Total**	20	40%	5.05	0.70	3.6 – 5.9
	Right ear OME	12	24%			
	left ear OME	8	16%			
Bilateral	30		60%	4.90	0.74	3.5 - 6
Total	50		100%	4.96	0.72	3.5 - 6

The average SRTs of the right and left ear, as well as the average PTA-HL are shown in Table 7.3, where it can be demonstrated that the average SRTs in bilateral OME (42.99 dBA ± 7.19 dBA) were higher than those in unliteral OME (38.50 ± 6.2 dBA), but the difference was not significant (Independent t-test: $p = 0.965$).

Table 7.3 Average PTA-HL and SRT (right and left ears) mean, SD, minimum and maximum in unilateral and bilateral OME

	Variable	n	M	SD	Min	Max
Average PTA-HL (dB HL)						
Unilateral	**Average (Rt and Lt)**	20 participants	16.55	6.20	4.00	28.00
	Ears with OME	20 ears	17.99	8.37	6.00	36.00
	Ears with No OME	20 ears	15.11	5.01	2.00	23.00
Bilateral	**Average (RT and Lt)**	30 participants	21.17	7.19	10.00	39.50
Average SRT (dB A)						
Unilateral	**Average (Rt and Lt)**	20 participants	38.50	8.18	26.63	56.83
	Ears with OME	19 ears	40.88	8.6	27.00	59
	Ears with No OME	18 ears	35.19	8.88	24.33	54.66
Bilateral	**Average (RT and Lt)**	30 participants	42.99	8.53	29.47	59.00

A boxplot of the average SRTs by type of OME (unilateral and bilateral), and a boxplot of the SRT by OME status (OME and no OME) in participants with unilateral OME are displayed in Figure 7.4 and Figure 7.5, respectively.

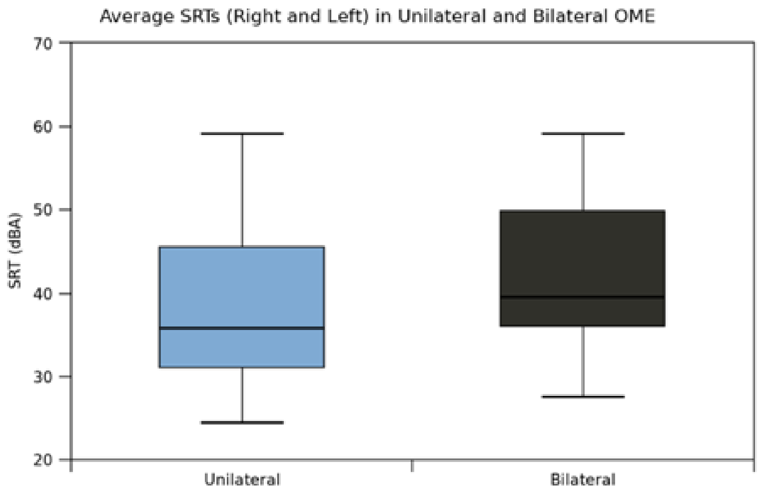


Figure 7.4 Boxplot of average (right and left) SRTs (dBA) in unilateral (n=20 participants) and Bilateral (n=30 participants) OME

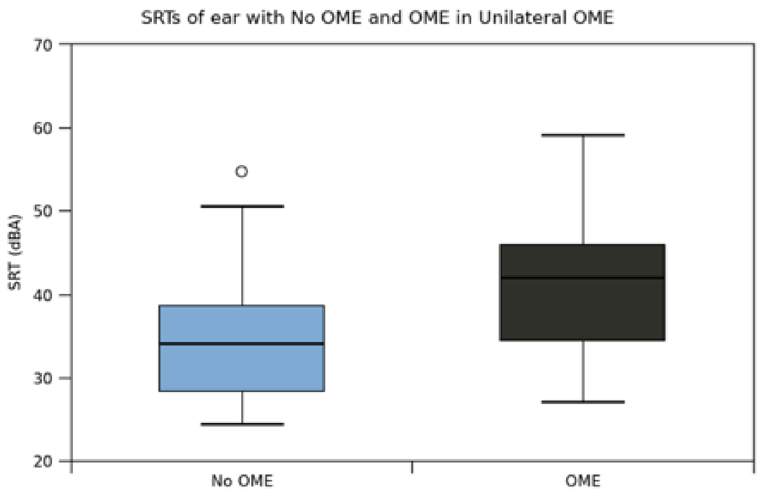


Figure 7.5 Boxplot of SRTs (dBA) in participants with unilateral OME by OME status (No OME: n = 19 ears, OME = 18 ears)

Effect of OME status, laterality, and age on SRTs

A linear mixed model was performed to evaluate the relationship between OME condition (status and laterality) and SRTs, controlling for subject age. The dataset contained SRT measurements for 50 participants, typically for both ears.

The analysis labelled each measurement with a condition, coded as a factor with three levels: OME unilateral, No OME unilateral and OME bilateral. In the LMM analysis that followed, the unilateral OME condition was taken as a baseline, so that the estimated parameters correspond to the comparisons of interest.

Exploratory analysis

Within each group, the effect of age in years was seen to be roughly linear (Figure 7.6). Combining the three OME groups, the overall marginal effect of year on SRT was linear as well (Figure 7.7).

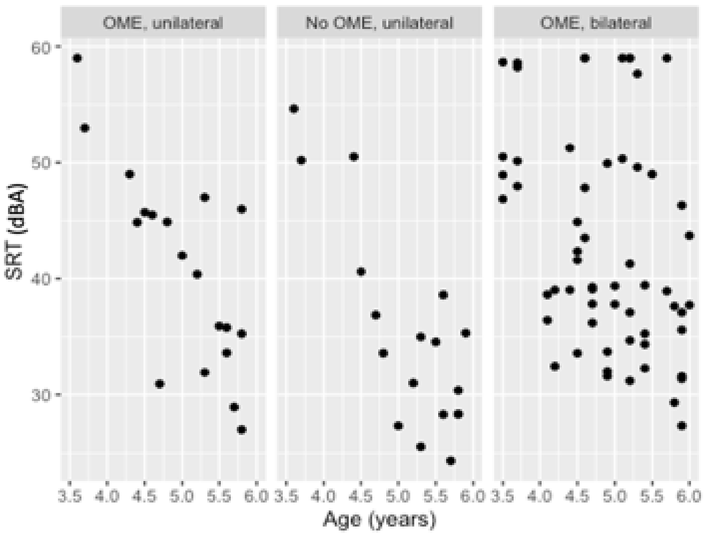


Figure 7.6 Exploratory scatterplots of SRT (dBA) and age in (1) OME, unilateral: OME ears in children with unilateral OME (n = 19 ears), (2) No OME, unilateral: No OME ears in children with unilateral OME (n = 18 ears), and (3) OME, bilateral: OME ears in children with bilateral OME (n = 60 ears)

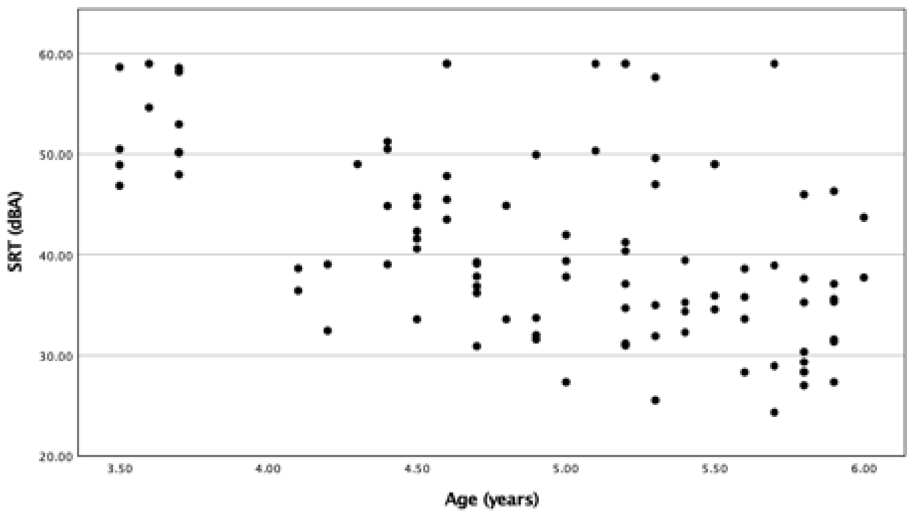


Figure 7.7 Exploratory scatterplots of the SRTs (n = 97 ears) and age (years)

The overall marginal effect of OME condition is shown in the boxplot below. The OME conditions appeared to be broadly similar to each other, whereas the distribution of SRT in the No OME condition was shifted lower relative to the OME conditions (Figure 7.8).

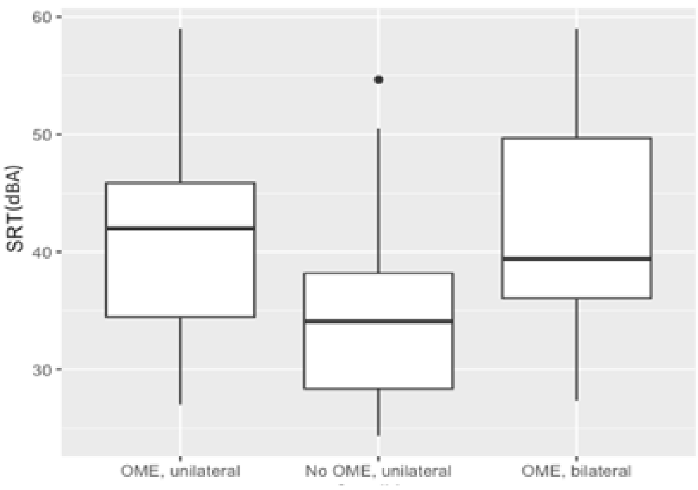


Figure 7.8 Boxplot of SRT (dBA) by OME condition (1) OME, Unilateral: OME ears in participants with unilateral OME (n = 19 ears), (2) No OME, unilateral: No OME ears in participants with Bilateral OME (n = 18 ears), and (3) OME, bilateral: OME ears in participants with bilateral OME (n = 60 ears)

The age distributions for participants' ears with unilateral and bilateral OME were broadly comparable, as seen in the boxplot Figure 7.9.

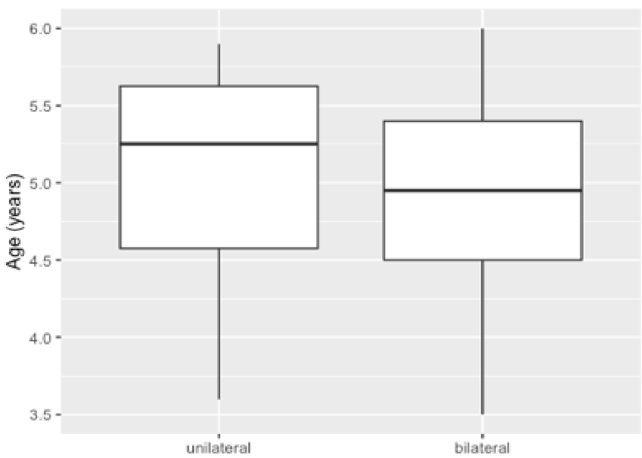


Figure 7.9 Boxplot of the age (years) by OME laterality (1) Unilateral (n = 37 ears), (2) Bilateral (n = 60 ears)

Mixed model analysis

The Q-Q plot and residuals from the model appeared to be normally distributed, suggesting no obvious difficulties with the error structure of the model, and that normality and homoscedasticity assumptions were met.

The model assumed the following form for y_{ij}, the SRT measurement for ear i (i = 1 for left ear, i = 2 for right ear) from subject j.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + \beta_3 a_j + b_j + \epsilon_{ij},$$

where:

- x_{ij} is 1 if ear i in subject j does not have OME and 0 otherwise
- z_j is 1 if subject j has bilateral OME, and 0 otherwise,
- a_j is the age of subject j in years,
- $b_j \sim N(0, \sigma_b^2)$ is a random effect for subject j,
- $\epsilon_{ij} \sim N(0, \sigma^2)$ is random error.

In this model, β_1 represented the SRT difference between a No OME ear and an OME ear, in participants with unilateral OME, controlling for age. β_2 represented the SRT difference between an OME ear in participants with bilateral OME and an OME ear in participants with unilateral OME, controlling for age. β_3 represented the average change in SRT with every one year of age.

The model was fit using REML (restricted maximum likelihood), using bootstrap to compute standard errors by resampling existing data multiple times to generate empirical estimates of the sample distribution (Field, 2009). The model parameters are given in Table 7.4.

Table 7.4 Model parameter and output for the effect of OME conditions and age on SRTs

Comparisons	**Beta coefficients**	**p-value**	**95%CI (dBA)**	
			Lower	**Upper**
Effect of OME status, laterality on SRTs, and age on SRT				
OME vs. No OME (Uni) (β_1)	-4.96	0.006*	-8.68	-1.78
OME (Uni) vs. OME (Bil) (β_2)	1.26	0.572	-3.47	5.68
Change with age (years) (β_3)	-6.35	<0.001*	-9.38	-3.70
Effect of OME status and age on SRT				
OME vs. No OME (Uni) (β_1)	-5.347	.002*	-8.602	-2.093
Change with age (years) (β_3)	-6.418	<0.001*	-9.184	-3.652
Effect of OME laterality and age on SRT				
OME (Uni) vs. OME (Bil) (β_2)	-3.629	.082	-7.721	0.463
Change with age (years) (β_3)	-6.507	<0.001*	-9.305	-3.708

*(Age is a covariate, OME: OME status, No OME: No OME status, Uni: Unilateral, Bil: Bilateral, *significant p-value < 0.05)*

Controlling for age, the SRT measurement in ears without OME was 4.96 dB (95%CI 1.78 to 8.68 dB) lower than in ears with OME in participants with unilateral OME. The SRT measurement in participants with bilateral OME was not significantly different from the SRTs in OME ears in participants with unilateral OME, after controlling for OME status and age. The results suggested that the SRT in an ear does depend on its OME status but does not significantly depend on the OME status of the contralateral ear. The analysis also indicated that there was a highly significant effect of age, with an estimated improvement of 6.35 dB per year.

<u>Objective 2 To measure the ability of SRTs to predict PTA thresholds</u>

A LMM was performed to evaluate the relationship between PTA-HL and SRTs, taking into consideration OME condition (status and laterality) and controlling for participants' age. The dataset contained PTA-HL and SRT measurements for 50 participants, typically for both ears.

<u>Exploratory analysis</u>

Assessing by the scatterplots of the relationship between SRT and PTA-HL grouped by OME status (Figure 7.10) and OME laterality (Figure 7.11), the relationship between SRT (dBA) and PTA-HL (dBHL) was seen to be linear.

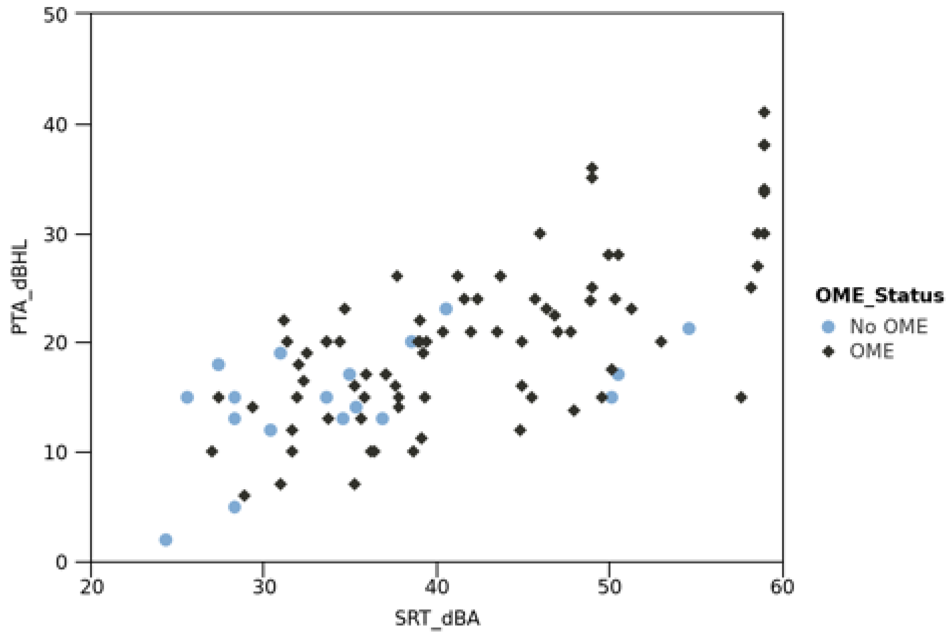


Figure 7.10 Scatterplot of SRT (dBA) and PTA-HL (dBHL) grouped by OME Status (No OME: n = 19 ears, OME: n = 78 ears)

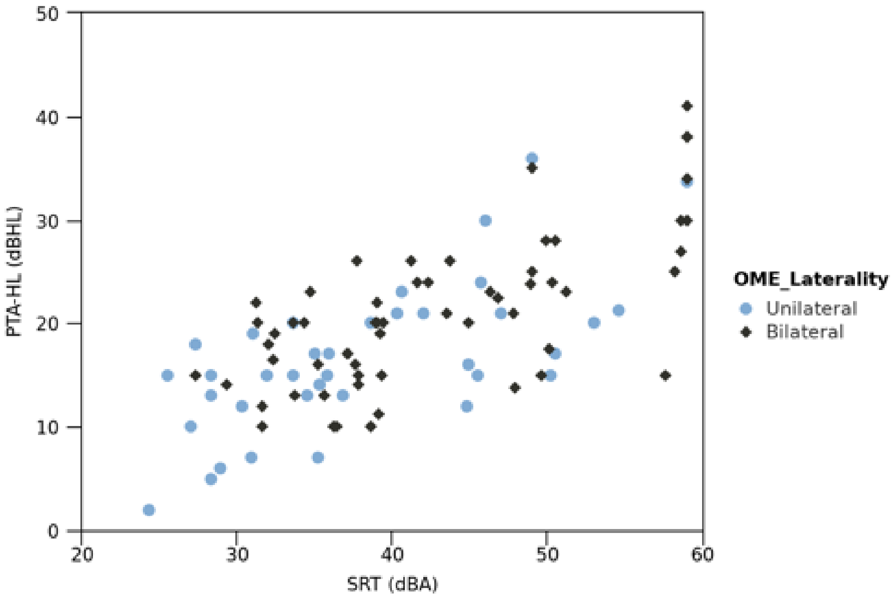


Figure 7.11 Scatterplot of SRT (dBA) and PTA-HL (dBHL) grouped by OME Laterality (Unilateral: n = 37 ears, Bilateral: n = 60 ears)

Mixed model analysis

The Q-Q plot and the residuals from the model appeared to be normally distributed, suggesting no obvious difficulties with the error structure of the model, and that normality and homoscedasticity assumptions were met.

The model assumed the following form for y_{ij}, the PTA-HL measurement for ear i (i = 1 for left ear, i = 2 for right ear) from subject j.

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \beta_2 z_j + \beta_3 a_j + \beta_4 s_j + b_j + \epsilon_{ij}$$

where:

- x_{ij} is 1 if ear i in subject j does not have OME and 0 otherwise
- z_j is 1 if subject j has bilateral OME, and 0 otherwise,
- a_j is the age of subject j in years,
- s_{ij} is the SRT of side i in subject j in dB A,
- $b_j \sim N\left(0, \sigma_b^2\right)$ is a random effect for subject j,
- $\epsilon_{ij} \sim N(0, \sigma^2)$ is random error.

In this model, β_1 represented the PTA-HL difference between a No OME ear and an OME ear, in participants with unilateral OME, controlling for SRT. β_2 represented the PTA-HL difference between an OME ear in participants with bilateral OME and an OME ear in participants with unilateral OME, controlling for SRT. β_3 represented the average change in PTA-HL with age of 1 year. β_4 represented the average change in PTA-HL (dBHL) with one-unit SRT (dBA).

The model was fit using REML, using bootstrap to compute standard errors by resampling existing data multiple times to generate empirical estimates of the sample distribution (Field, 2009). The model parameters are given in Table 7.5.

Table 7.5 Model parameter and output for the effect of SRT, age and OME conditions on PTA-HL

Comparisons	**Beta coefficients**	**p-value**	**95%CI (dB HL)**	
			Lower	**Upper**
Effect of OME status and SRT on PTA-HL (dBHL)				
OME vs. No OME (β_1)	-0.214	0.837	-2.275	1.846
OME (uni) vs. OME (Bil) (β_2)	-1.927	0.220	-5.027	1.172
Change with SRT (dB A) (β_3)	0.494	<0.001*	0.374	0.613
Effect of age and SRT on PTA-HL (dBHL)				
Change with age (years) (β_3)	.951	.387	-1.222	3.125
Change with SRT (dB A) (β_4)	.531	<0.001*	.410	.652
Effect of SRT on PTA-HL (dBHL)				
Change with SRT (dB A) (β_3) Intercept (β_0) = -1.510	0.510	<0.001*	0.398	0.621

*(Covariates: age (years) and SRT (dBA), OME: OME status, No OME: No OME status, Uni: Unilateral, Bil: Bilateral, *significant p-value < 0.05)*

The model showed, when adding OME status and laterality to the model and controlling for SRTs, that there was a significant increase of 0.49 dBHL of PTA-HL with every unit dBA increase in SRT ($p < 0.001$).

Given there was no significant effect of OME condition on PTA over and above that which is explained by SRTs, therefore the OME conditions predictors were removed when assessing the ability of SRTs to predict PTA-HL.

The statistical significance of adding the interaction term β_3 (age) to the model was assessed using the change in "-2log(likelihood)" for the model with and without β_3 (Twisk, 2019), which showed no significant effect of age on the relationship between PTA-HL and SRT (Appendix H.1).

In order to obtain the prediction formula of PTA from SRTs, the LMM was conducted with SRT (β_4) only as a fixed effect, where the result showed a significant increase of 0.51 dB-HL in PTA-HL with every unit increase in SRT ($p < 0.001$), and an intercept (β_0), the value of dependent (PTA-HL) when the independent (SRTs) = 0 dBA (Twisk, 2019), of -1.51 dBHL.

$$y_{ij} = \beta_0 + \beta_3 s_{ij}$$

PTA-HL = -1.51 + (0.51 x SRT)

The relationship between SRTs and different frequency combination averages of PTA-HL was also assessed. Pearson correlation was conducted between PTA-HL (average right and left) and SRT average (average right and left) for all participants in the study, including those from the pilot (n = 56). This analysis was done for PTA-HL average of five frequency combinations:

1- PTA average 0.5 and 1 kHz (dB HL)
2- PTA average 1 and 4 kHz (dB HL)
3- PTA average 0.5, 1, and 4 kHz (dB HL)
4- PTA average 0.5, 1, 2 and 4 kHz (dB HL)
5- PTA average 0.25, 0.5, 1, 2 and 4 kHz (dB HL)

Visual inspection of the scatterplots indicated a linear relationship between PTA-HL of different combinations and SRT (Figure 7.12). The SREs were all within ±3 indicating that there are no significant outliers.

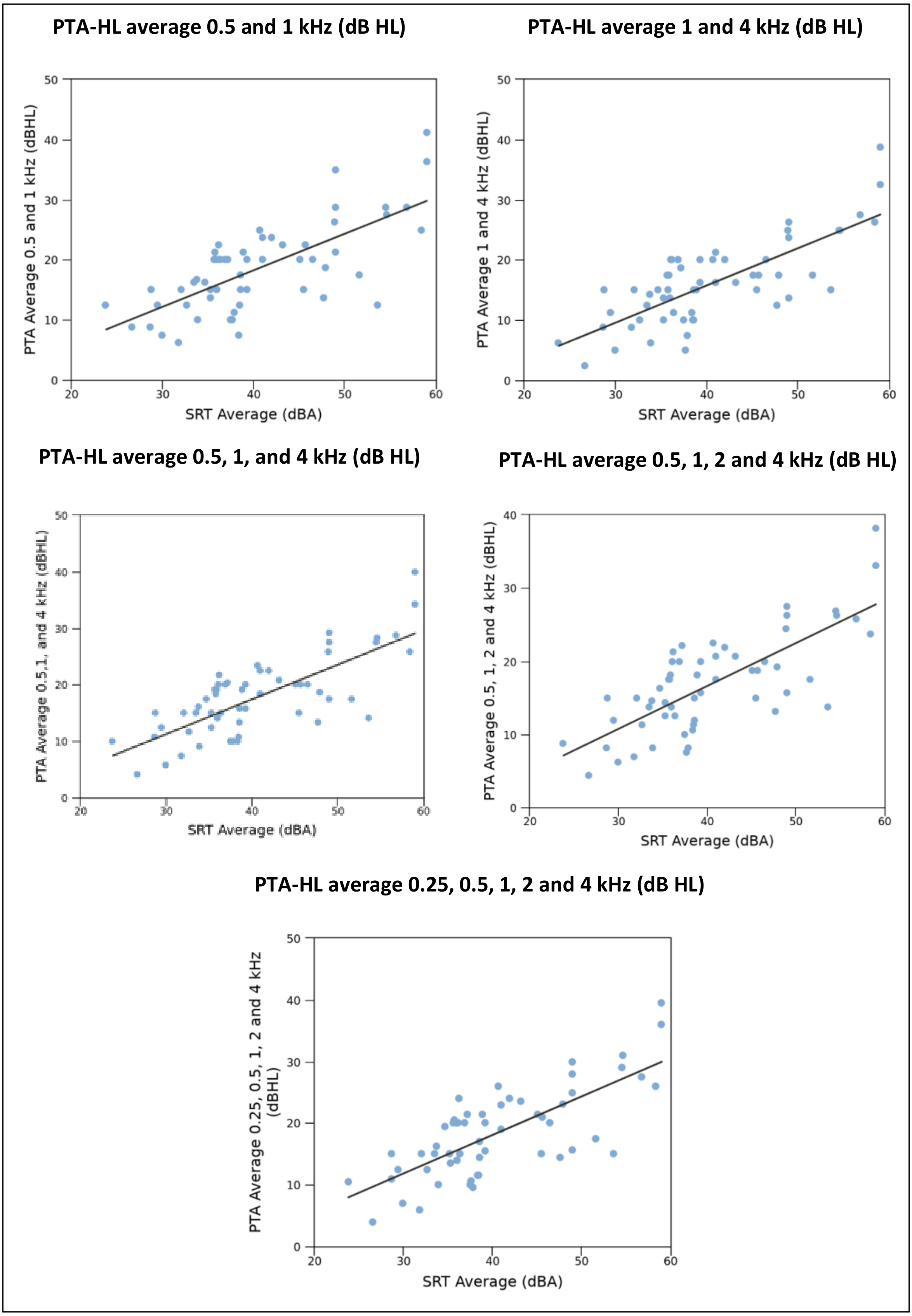


Figure 7.12 Scatterplots with fitted regression line of the different PTA-HL average of frequency combinations and SRTs

All frequency combinations yielded a statistically significant (p <0.001) positive relationship between SRTs and PTA-HL with similarly high correlation coefficients (*r* > 0.7) indicating a large effect size (Table 7.6). The 95%CI of correlations (*r*) of the relationships were overlapping, indicating no significant differences between the correlations of these relationships.

Table 7.6 Pearson correlation of the relationship between Average (right and left) SRTs and PTA-HL for all participant (n=56) in five different frequency combinations

Frequency combination	**Pearson Correlation (*r*)**	**95%CI**		**Significance *p*-value**
		Upper	**Lower**	
0.5 and 1	0.71	0.56	0.82	<0.001
1 and 4	0.77	0.63	0.86	<0.001
0.5, 1 and 4	0.75	0.62	0.85	<0.001
0.5, 1, 2 and 4	0.75	0.60	0.84	<0.001
0.25, 0.5, 1, 2, and 4	0.74	0.60	0.84	<0.001

Objective 3 To measure the effect of OME on the AOM6

The AOM6 (Total) for all participants was 2.76 (SD = 1.06), and the AOM6 (HaS) was 2.6 (SD = 1.3). Figure 7.13 displays boxplots of the of the scores of AOM6 (Total) and AOM6 (HaS) based on OME type (unilateral or bilateral)

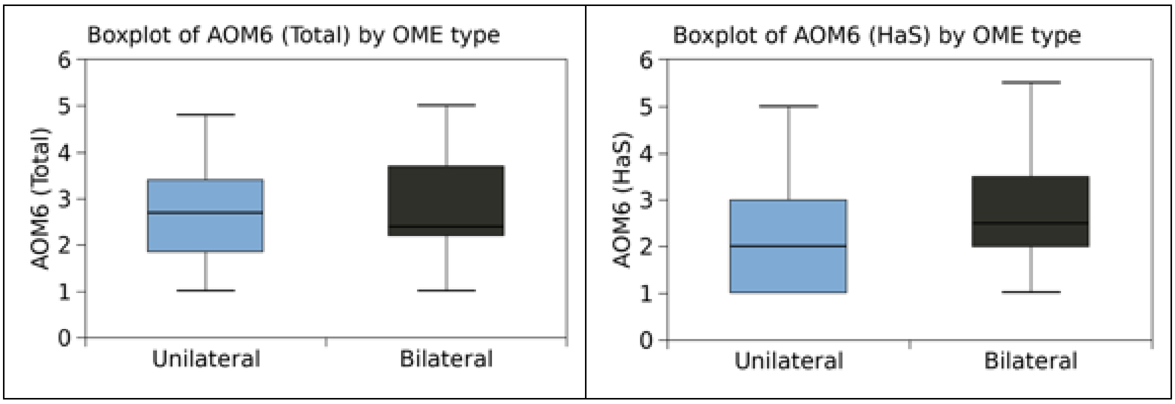


Figure 7.13 Boxplots of the scores AOM6 (Total) and AOM6 (HaS) by OME type (unilateral n= 20 or bilateral n= 30)

An independent samples t-test on the difference in AOM6 (Total) between unilateral and bilateral OME was applied, normality based on the Shapiro-Wilk test was assumed for both groups (p > 0.05). There was no significant difference in AOM6 (Total) scores between unilateral and bilateral

OME (F (1, 48) = 0.28, p = .601). A similar analysis was performed for the AOM6 (HaS), and even though the normality assumption was violated for the unilateral group, the decision to proceed with the independent t-test was made because of the robustness of t-tests, as well as the presence of normality assumption with the other combinations. There was also no significant difference in the AOM6 (HaS) between unilateral and bilateral OME (F(1, 48) = 2.48, p = .122). Table 7.7 displays the mean, SD and p-value from each AOM6 score in unilateral and bilateral OME.

Table 7.7 AOM6 (Total) and AOM6 (HaS) in overall participants, as well as in participants with unilateral and bilateral OME

	Overall		**Unilateral OME**		**Bilateral OME**		***p-value***
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	
AOM6 (Total)	2.76	1.06	2.67	1.03	2.83	1.08	0.601
AOM6 (HaS)	2.6	1.3	2.25	1.25	2.83	1.3	0.122

Regarding the hearing item in the AOM6 questionnaire, the most common answer reported by parents of participants with unilateral and bilateral OME was “Not present/ No problem” (40% and 30%, respectively). The second most common answer reported by parents of participants with unilateral OME was “Hardly a problem at all” (20%), whereas, in children with bilateral OME, it was “Somewhat of a problem” (26.7%). The least common answer in children with bilateral OME was “Very much a problem” (3.3%) and “Extreme problem” (3.3%), whereas in participants with unilateral OME, none of the parents reported these two answers (0%), and their least common answer was “Quite a bit of a problem” (10%) (Figure 7.14).

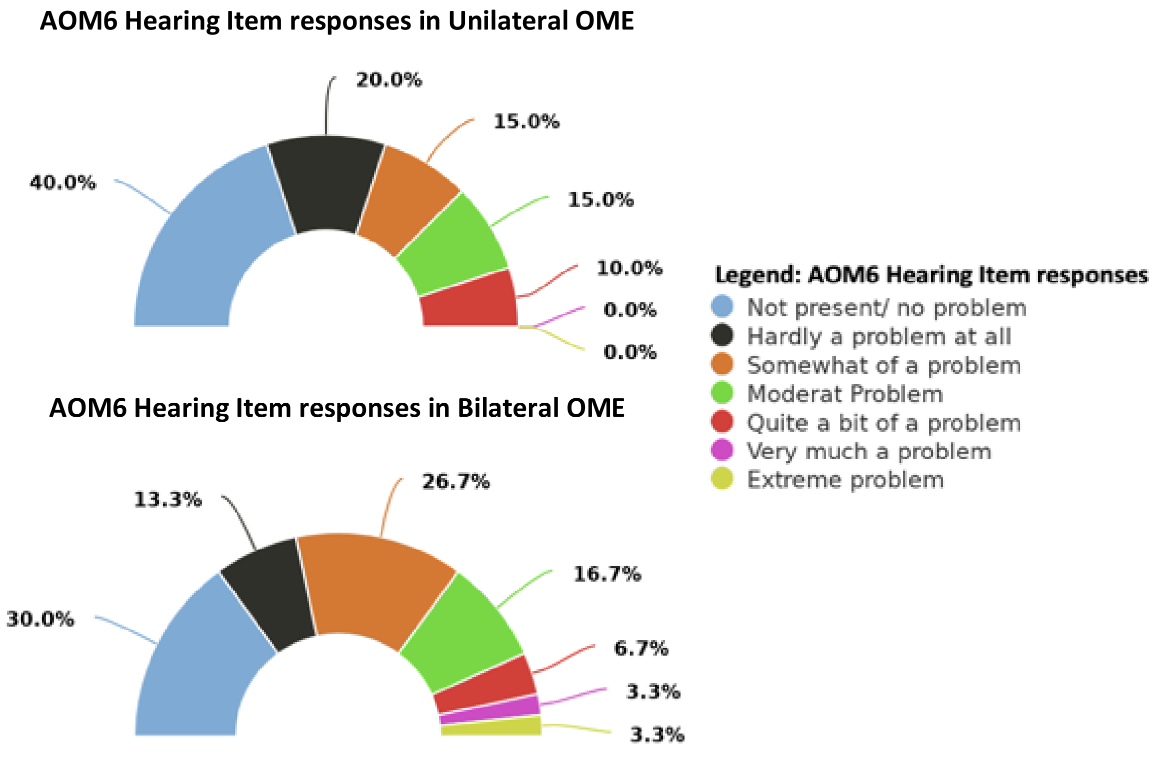


Figure 7.14 Chart of the percentages of responses on the hearing item on the AOM6 in Unilateral and bilateral OME

7.5.2 Aim 2 To investigate the relationship between speech recognition and QoL

Objective To measure the relationship between the AOM6 and SRT and PTA-HL

Pearson correlation was performed to assess the relationship between SRT and AOM6 (Total) and AOM6 (HaS), as well as PTA-HL and AOM6 (Total) and AOM6 (HaS). The AOM6 scores were obtained from the first repeat. The assumption of linearity was met for all combinations (Figure 7.15).

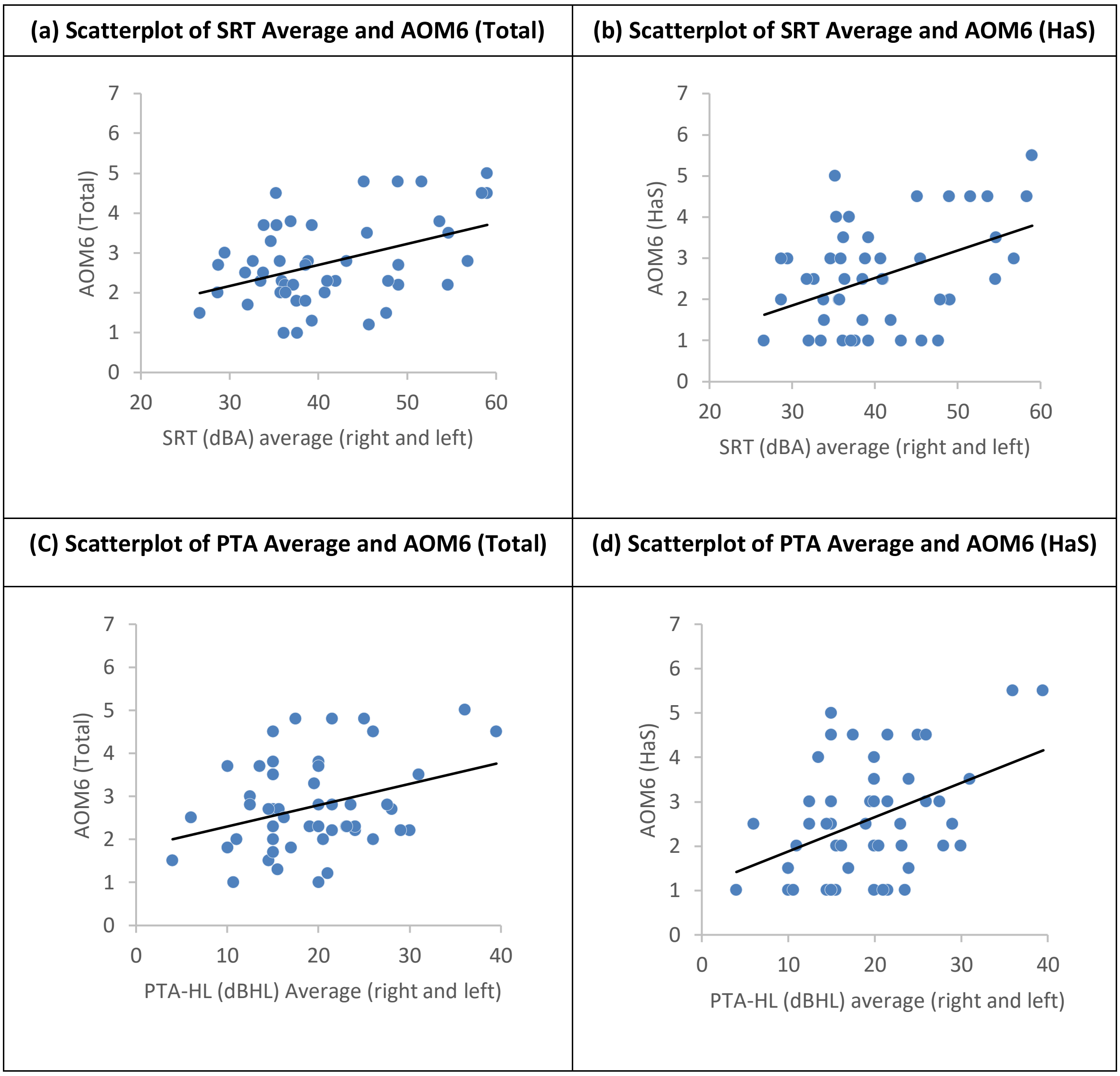


Figure 7.15 Scatterplots with regression line added for the combinations (n= 50) (a) SRT and AOM6 (Total), (b) SRT and AOM6 (HAS), (c) PTA-HL and AOM6 (Total), and (d) PTA-HL and AOM6 (HAS).

The result of the correlation was examined based on an alpha value of 0.05. A significant positive correlation was observed for all combinations, with a correlation ranging between 0.33 and 0.43 for all combinations indicating a moderate effect size. This correlation suggested that as SRT and PTA-HL averages increase, AOM6 (Total) and AOM6 (HAS) scores tend to increase. The results of the correlations are displayed in Table 7.8. Further analysis to assess the significance of the difference between the two correlation coefficients of AOM6 (Total) and SRTs, and AOM6 (Total) and PTA showed no significant difference between the correlations ($p = 0.15$) (Appendix H.2).

Table 7.8 Pearson Correlation Results Between SRT and AOM6, SRT and AOM6 (HAS), PTA-HL and AOM6, and PTA-HL and AOM6 (HAS)

Combinations		***r***	**n**	**95%CI**		**p-value**
				Lower	**Upper**	
SRT	**AOM6 (Total) score**	.43	50	.17	.63	.002
	AOM6 (HAS) score	.44	50	.19	.64	.001
PTA-HL	**AOM6 (Total) score**	.33	50	.06	.56	.018
	AOM6 (HAS) score	.42	50	.16	.63	.002

A Spearman's rank-order correlation assessed the relationship between the AOM6(HI) and SRTs. All fifty participants were assessed. Preliminary analysis showed the relationship to be monotonic, as assessed by visual inspection of the scatterplots in Figure 7.16. Table 7.9 displays the spearman rank correlation between the AOM6(HI) and SRTs and PTA-HL. There was a weak positive significant correlation between Average PTA-HL and the AOM6(HI), whereas the SRTs did not correlate significantly with AOM6(HI).

(1) Average SRTs and AOM6 Hearing item

(2) Average PTAs and AOM6 Hearing item

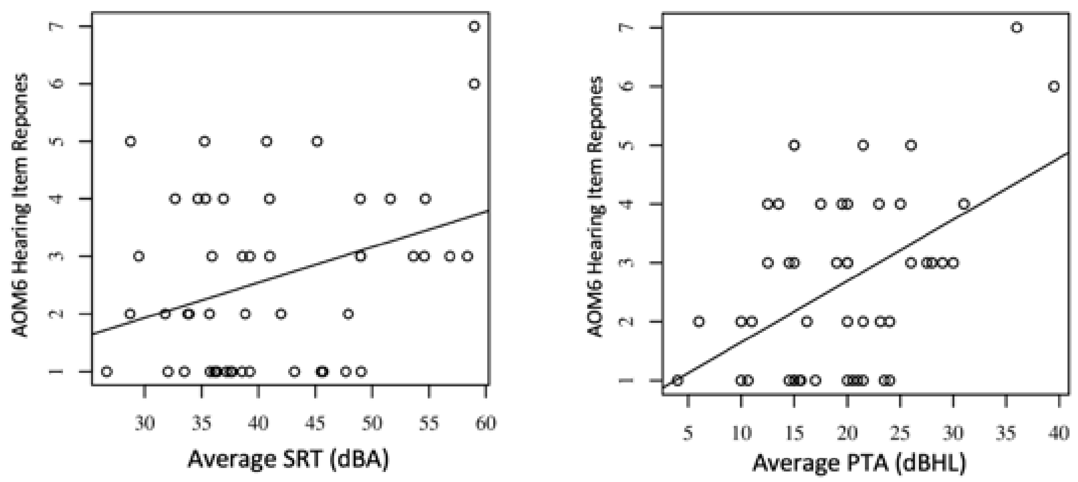


Figure 7.16 Scatter plot of the relationship between Average (right and left) (1) SRTs and (2) PTA and the AOM6 Hearing item responses (each n = 50)

Table 7.9 Spearman correlation between the SRTs and PTA-HL (independent variable) and AOM6 (HI) (dependent variables)

Combination	***n***	**Spearman's rho Correlation Coefficient**	**95%CI**		**P-value**
			Lower	**Upper**	
AOM6(HI) and SRT Average (dBA)	50	.23	-0.06	0.47	.114
AOM6(HI) and PTA Average (dBHL)	50	.34	0.06	0.56	.017

7.5.3 Aim 3 To assess the reliability of the of the PAAST SiQ and AOM6

Objective 1 To assess the test-retest reliability of the PAAST SiQ

Exploratory plots

There were two measurements for each of the 32 subjects. A scatterplot of these measurements is displayed in Figure 7.17. The data included three measurements outside the specified age range for the study, which was 3.5 to 6 years, these measurements (coloured red in the plot below) were from participants from the pilot study and did not appear to be atypical. From this plot, there was a strong linear relationship between the two measurements.

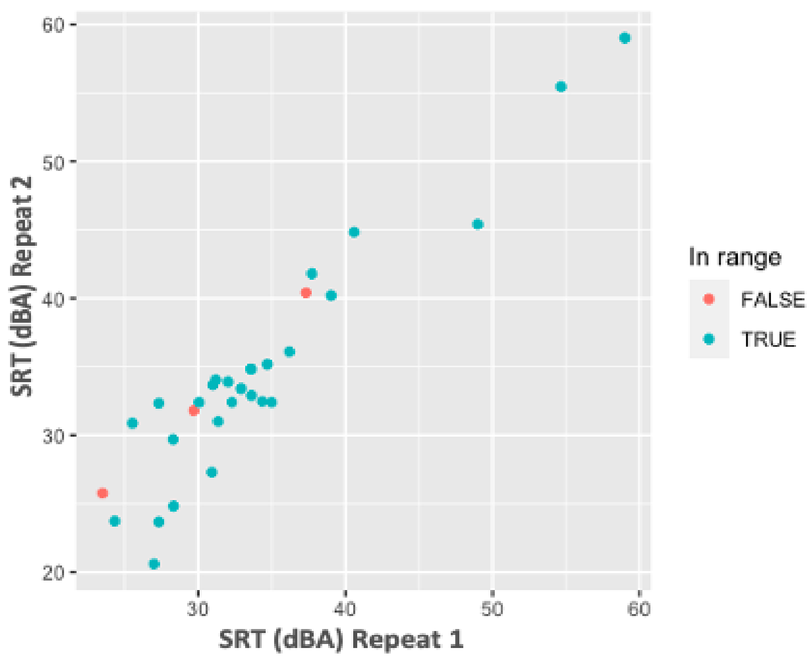


Figure 7.17 Exploratory scatter plot of the relationship between the first and the second SRT (dBA) measurement for each participant (n = 32). False represents participants from the pilot who were outside the age range of the study.

A RM-ANOVA was conducted to determine whether there were systematic variations in the two repeats (Repeat 1 and 2) of SRTs. Although there were outliers in the raw data, the SREs were all within the range of ±3 indicating that there were no significant outliers. The data was not normally distributed at either time point, as assessed by Shapiro-Wilk test ($p < .05$), but since the sample size was not small (n=31), the levels of the within-subjects factor are similarly skewed, and the RM-ANOVA was considered a robust test to non-normality, the decision has been made to proceed with RM-ANOVA. The assumption of sphericity was not tested because there were only two repeated measurements. The mean SRT in Repeat 1 was 34.25 (SD = 8.81) dBA, 34.86 (SD = 9.3) dBA in Repeat 2. The repeats did not elicit statistically significant changes in SRTs over time ($F(1, 30) = 1.629$, $p = .212$, partial $\omega 2 = .052$) with a nonsignificant increase of 0.62 dB (95%CI: -0.37, 1.61 dB) in Repeat 2 (Figure 7.18).

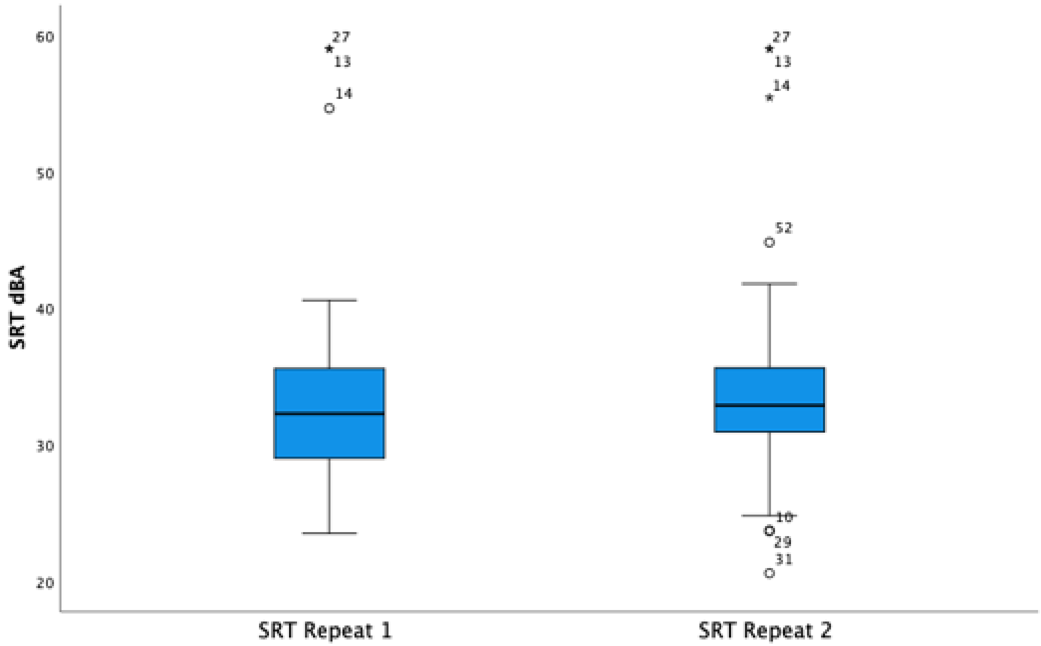


Figure 7.18 Average SRTs in PAAST SiQ in Repeat 1 and 2 (n= 32 in each boxplot)

Bland-Altman plot

The plot below showed the difference in the two measurements, plotted against their average (Figure 7.19). The mean difference is shown as a dashed red line. The green dashed lines are at the mean difference ± 2 standard deviations. Assuming the difference to be normally distributed, $\sim 95\%$ of observations should lie within the two lines. For these data, 97% lied within the two lines, and there were no more observations outside the lines than would be expected by chance. This finding suggested good reliability.

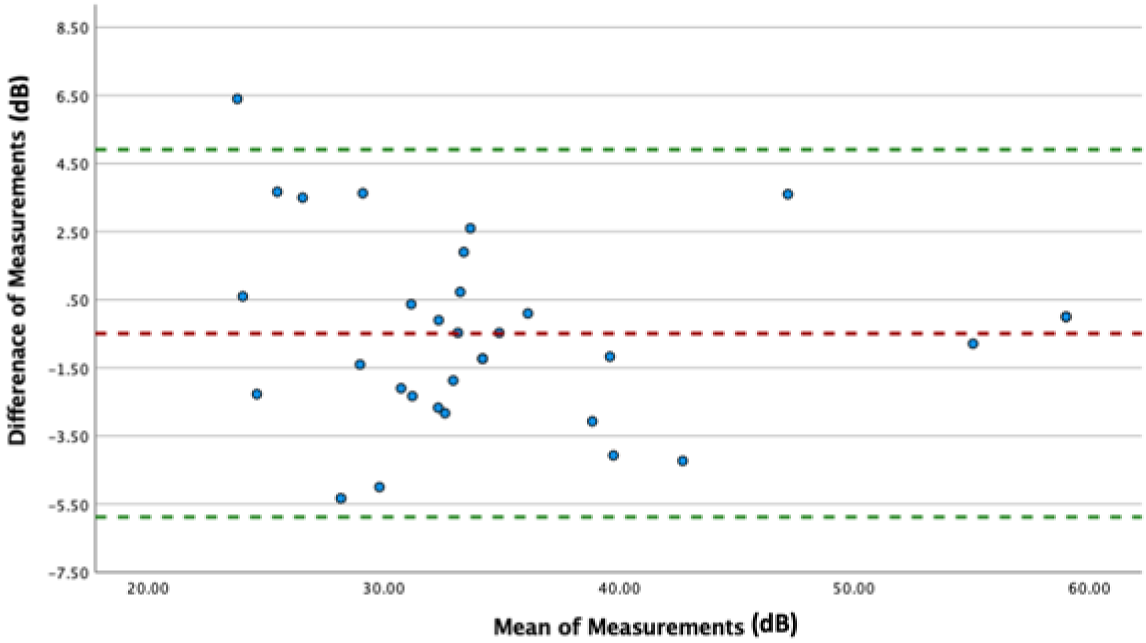


Figure 7.19 Bland-Altman Plot difference in the two measurements of SRT (R1 and R2), plotted against their average.

Variability and stability

Variability of the test was measured by the within-subject standard deviation (SDω), which is calculated using the formula proposed by (Bland and Altman, 1996),

$$\sqrt{(\sum_i^n d^2)/2n},$$

where d is the difference between the scores obtained by the i[th] subject, and n is the number of subjects (participants) (Bland and Altman, 1996). The SDω was 1.95 (95%CI [1.18, 2.71 dB]).

Repeatability of the scores is the minimum value that would inform a statistically significant difference between any two measurements. The repeatability was measured by multiplying the SDω by √2 x 1.96 (Bland and Altman, 1996). The repeatability of the PAAST was 5.39 dB.

The model

The one-way random effects model contained a random effect for each individual, as well as an overall mean (Liljequist, *et al*., 2019). The model assumed the following form for y_{ij}, the ith measurement on individual j, where μ is a constant, while $\mu + b_j$ is the "true score" for subject i.

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

For the data here, the ICC in the one-way random effects model was 0.96. This value represented excellent test-retest reliability, which was consistent with the strong linear relationship seen in the scatter plot of first against second measurement.

Objective 2 To assess the test-retest reliability of the AOM6

A one-way RM-ANOVA was conducted to determine whether there was a systematic variation between the AOM6 Sum-Total in the first repeat and the second repeat 7 days later. The SREs were all within the range of ±3, indicating that there were no outliers. The data was normally distributed at both time points, as assessed by Shapiro-Wilk test (Repeat 1: $p = 0.220$, Repeat 2: $p = 0.163$). The AOM6 (Total) in Repeat 1 was 2.83 (SD = 1.3), and 3.00 (SD = 1.45) in Repeat 2 (Figure 7.20). The repeats did not elicit statistically significant changes in scores over time ($F(1, 27) = 3.05$, p = .433, partial ω2 = 0.02) with a nonsignificant increase of 0.168 (95%CI: -0.27, 0.60) in R2 scores.

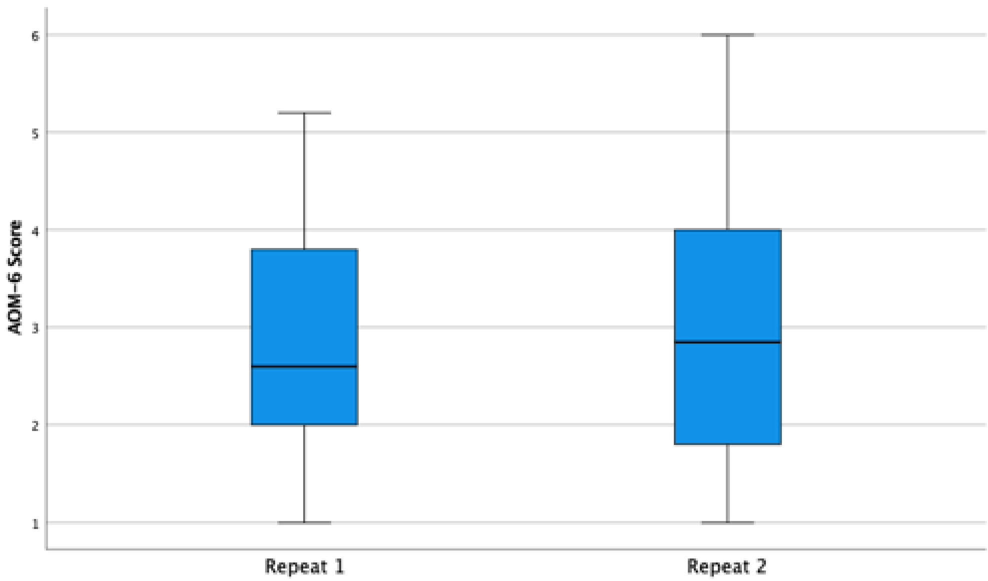


Figure 7.20 Boxplot of AOM6 score in Repeat 1 and 2 (n= 28 in each boxplot)

Bland-Altman plot

The plot below (Figure 7.21) shows the difference in the two measurements, plotted against their average. The mean difference is shown as a dashed red line. The green dashed lines are at the mean difference ± 2 standard deviations. Assuming the difference to be normally distributed, $\sim$ 95% of observations should lie within the two lines. For these data, 93% lied within the two lines.

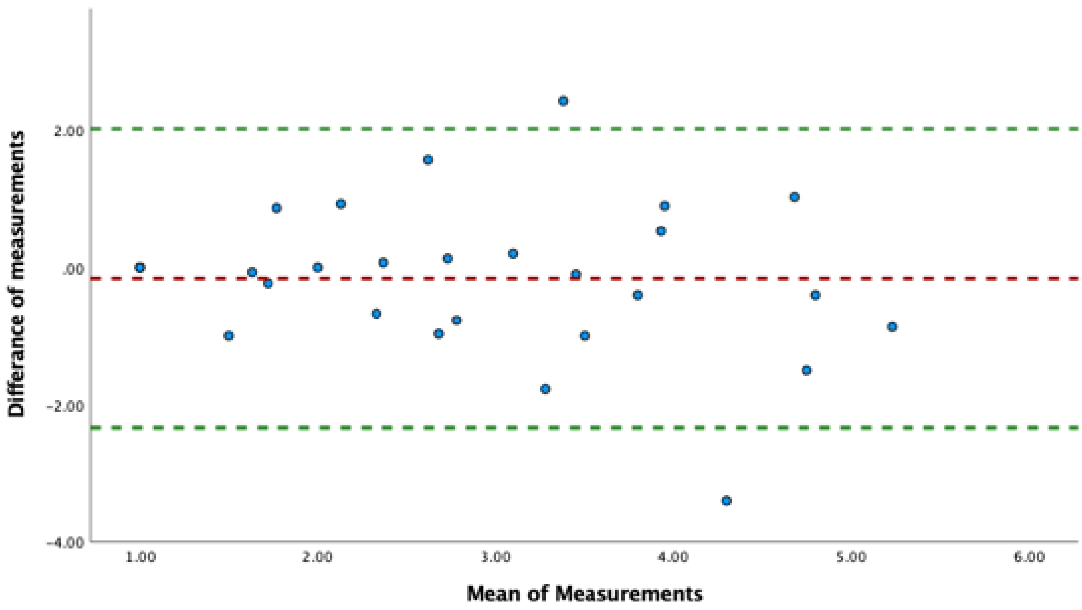


Figure 7.21 Bland-Altman Plot difference in the two measurements of AOM6 (R1 and R2), plotted against their average (n = 28)

The Model

the ICC in the one-way random effects model was 0.67, which represented a moderate test-retest reliability.

7.6 Discussion

Outcome measures are important in deciding on the management plan in any condition, including OME, where watchful waiting is an option that could be implemented and, in some cases, extended to 6 months prior to surgical or medical interventions (Rosenfeld *et al.*, 2022). This is especially important knowing that 81.8% of parents in Saudi Arabia were willing to follow the watchful waiting approach in cases where they were asked to participate in the decision plan (Alsuhaibani *et al.*, 2020). This notion sheds a light on the importance of having an appropriate set of outcome measures not only in the clinical setting, but also in research, where the effectiveness of treatment of OME must be evaluated with the right set of outcome measures. As important as clinical examination and impedance tests are as outcome measures, it has been increasingly acknowledged that assessment of hearing and QoL should be included in the evaluation of children with OME (Homøe *et al.*, 2020; Rosenfeld *et al.*, 2022). In this study, speech recognition in quiet was researched as a measure of hearing, alongside the QoL.

To the researcher's knowledge, no published studies looked at speech recognition and QoL in children with OME in the Arab world or Saudi Arabia.

This chapter discussed the final phase of this project, ensuring speech recognition and QoL using the PAAST SIQ and AOM6, respectively, could be considered as good outcome measures in Arabic-speaking children with OME. The aim of this experiment was to ensure the validity and reliability of both tools in Arabic children with OME. This aim was achieved through a set of objectives, each objective is discussed individually in the following sections.

All children with OME who were approached in the study (except for two children who did not perform PTA) were able to perform PTA in at least three frequencies (0.5, 1, and 2 kHz) (n = 53). Most children (n = 48) were able to perform the PTA at frequencies 0.125 to 4 kHz. Five children (10%) were unable to perform the PAAST SiQ due to factors related to shyness and inability to recognise the pictures, especially in younger children (age range 3.4 – 4 years). These factors could be overcome by administering the test more appealingly in an application form on a smart device. Regarding the ability to recognise pictures, this may require training with the child to improve their vocabulary, which might have been affected by the CHL caused by OME. A community validation study of the MTT by Harries and Williamson (2000) proposed similar reasons to why three year old children might in some cases not perform the MTT, in addition to possible developmental delay or behavioural issues. In their study, 65 children were tested, 22% of which were unable to perform both standard tests, namely PTA, and MTT, and 3% were unable to perform the MTT (Harries and Williamson, 2000). These results were similar to those reported by Summerfield *et al.* (1994), where 5% of n=215 children aged two and older could not perform the test. In the study by Hall *et al*. (2007), which was a part of the ALSPAC large study, 8.4% children aged 2.5 years (out of n=1135) were unable to complete the AMTT, and 1.2% (out of n=1065) aged 3.5 years were also unable to perform the test. Another possible cause could be related to the impact of COVID-19 on children's development and behaviour, which is further discussed in the section on the effect of age on SRTs.

During the design phase of the current study, the scope of the thesis and the challenges presented by the COVID-19 pandemic required a flexible approach to methodology. Although the power calculation indicated a sample size of 64 would be ideal to achieve the aims of the current study, constraints necessitated a more feasible sample size of 50. This revised sample size aligns with previous studies, which have demonstrated variability in their sample sizes based on tailored research aims (Summerfield *et al*., 1994; Harries and Williamson, 2000; Hall *et al*., 2007). It is important to note that the primary focus of this research was on the significance of the results (p-value) rather than the magnitude of the effect size, which can be prone to overestimation in smaller samples (Slavin and Smith, 2009). This approach was intentional, aiming to establish preliminary evidence for a potential relationship between OME and SRT and QoL that future studies could explore in greater depth and with larger sample sizes, particularly as circumstances allow for a less restrictive research environment.

7.6.1 Aim 1 To investigate the effect of OME on the speech recognition and QoL

Objective 1 To measure the effect of OME status, laterality, and age on SRTs

One of the purposes of this study was to obtain data for speech recognition in children with OME and compare them to previous studies.

The mean of the average SRTs obtained from all n= 50 children using the PAAST SiQ was 38.5 dBA (SD = 8.18 dB) in unilateral OME and 42.22 dBA (SD = 8.53) in bilateral OME. The mean age of children was 5 years (SD = 0.7 years), and the age range was 3-6 years, which was somewhat similar to the range in the study by Hall *et al*. (2007). The similarity in age range allowed for comparison between both studies' data, where Hall *et al*. (2007) used the AMTT to obtain word recognition thresholds (WRTs), which are equivalent to SRTs in the current study. In their study, Hall *et al*. (2007) reported the mean SRT based on age groups, bearing in mind the difference in their method where they tested both ears simultaneously through loudspeakers. Hall *et al*. (2007) found that children aged 5 years had a mean SRT of 28.8 (SD = 4.3) dB and 38.5 (SD = 8.9) dB, in unilateral and bilateral OME, respectively. The results of the current study were more similar to those obtained by 2.5-year-olds in the study by Hall *et al.* (2007) (unilateral: 34.2 (6.0) dB, bilateral: 46 (8.7) dB). The difference in the SRTs between the two studies could be due to the difference in the test presentation, where it was monaural in the current study and binaural in the study by Hall *et al.* (2007), which could result in a difference of 1.7 dB, with better SRTs for binaural testing (Wilson, 2003). Another reason could be due to the negative effects of COVID-19 in children in the current study, who spent a large portion of their early years between lockdown and online-schooling, which might have affected their language acquisition. Details of the possible effects of COVID-19 preventative measures on SRTs are mentioned in the section discussing the relationship between SRTs and age.

Effect of OME status on SRTs

Otitis media can lead to CHL, evident in PTA thresholds (Cai and McPherson, 2016). According to a study examining auditory profiles of 146 ears of children aged 72 months to 153 months with OME, the mean PTA threshold was 26.8 dB HL with a rising configuration mostly affecting low frequency hearing thresholds with the lowest (best) thresholds at 2 kHz (Cai *et al.*, 2018). Other hearing abilities such as speech recognition can also be significantly worse in children with OME compared to their peers with no OME (Petinou *et al.*, 2001).

Based on this knowledge as well as data from five studies on the effect of OME on SRTs in quiet (Cai and McPherson, 2016), it was predicted that there would be a significant difference in SRTs in quiet in ears with OME compared to No OME. The results of the current study showed a significant difference in SRTs between OME and No OME status in unilateral OME (p = 0.001), with a 4.96 dB (95%CI 1.78 to 8.68 dB) mean elevation of SRT in OME ears.

A recent study by Brown *et al.* (2019) used the AMTT to measure speech recognition (tested with loudspeakers) while using a BC headset to investigate its use in school children suffering from fluctuating hearing loss due to OME or auditory processing disorder. Brown *et al.* (2019) found that in quiet, the median SRT without a headset was 20.5 dBA (range: 13-31 dBA), which was reduced to 11.5 dBA (range: 7-16 dBA) with the headset. The lower SRTs in their study and larger differences pre- and post-amplification (compared to the current study) may be attributed to the wider age range of children (4-11 years) and the method of measurement (loudspeakers) in their study.

Earlier studies also investigated the effect of OME on speech recognition using different methods to obtain SRTs. Sabo *et al.* (2003) assessed SRTs using conventional audiometry (i.e., speech stimuli presented in live voice through loudspeakers monaurally using audiometry or VRA), as well as speech awareness thresholds (SAR) in children aged 6 months to 36 months. They reported that in children aged 3.5 and 6 years with unilateral OME, the mean SRTs of OME ears (18.2, SD = 10.4 dB) were significantly higher (p = 0.05) than the mean SRTs of no OME ears (8.7, SD = 4.4 dB). The difference in mean SRT between OME and No OME ears was higher in the study by Sabo *et al.,* (2003) (approximately 10 dB) than that obtained in the current study (approximately 5 dB).

Interestingly, Sabo *et al.,* (2003) found that there was a significant difference (p < 0 .05) between the mean SRTs of no OME ears in children with unilateral OME and the mean SRTs (average across right and left ears) of age-matched children with no OME (mean = 6.2 dB). This difference was not observed in their study when comparing PTA thresholds, but there was still slightly better performance in the PTA in children with no OME compared to children with unilateral OME, suggesting that children with unilateral OME may suffer from residual inflammation, eustachian tube disfunction, or language delay due to periods of intermittent hearing loss.

Effect of laterality on SRT

Regarding the effect of laterality, no significant difference was found in SRTs between OME ears in bilateral and unilateral OME (p = 0.572). Sabo *et al.* (2003) reported similar findings to the current study, bearing in mind that the method they used was conventional speech audiometry. They found that in bilateral OME, the average (right and left ears) SRT (mean = 18.7, SD 10 dB) was slightly higher than the mean SRTs of OME ears of unilateral OME, but the difference was not significant. The finding in the current study suggested that the presence of OME in one ear does not affect the speech recognition in the contralateral ear, at least when tested monaurally.

One of the strengths of the current study was that speech recognition was assessed monaurally, which allowed for assessment of the effect of OME on each ear. This advantage was at the expense of assessing the effect of the presence of unilateral and bilateral OME on the SRTs when assessed binaurally, which could have allowed for comparison with studies that looked at such effect.

Effect of age on SRT

Development of hearing in children goes through several stages, and children aged 30 months (2.5 years) and above are expected to gradually develop the ability to recognise/identify words (Perigoe and Paterson, 2013). Considering the PAAST SiN was previously assessed for its usability in children aged 3 years and older (Al-Kahtani, 2020), it was predicted that SRTs could be assessed in children from the current study (age 3-6 years) using the PAAST SiQ. The fact that children in the current study were at an age range where speech recognition is still developing (Perigoe and Paterson, 2013) suggested the presence of subject variations in SRTs due to age. In the current study, age was controlled as a covariate in the LMM when looking at the effect of OME status and laterality on SRTs, and there was a significant ($p < 0.05$) 6.51 dBA SRT decrease (improvement) with every 1-year increase in age.

Studies varied in their findings on the relationship between SRTs and age. An early study by Jerger and Jerger (1982) (mentioned in Hall *et al*. (2007)) found a 4 to 6 dBA improvement in SRTs of monosyllabic words obtained at 50% correct at 3, 4, and 6 years. Summerfield *et al.,* (1994) found no significant improvement in SRTs with age using the AMTT in 215 children aged 2 – 13 years old. Hall *et al.* (2007) reported a 5 dB significant improvement in WRT using the AMTT as age increased from 31 to 61 months (2.5 – 5 years).

In a study that investigated the effect of hearing status on speech recognition in quiet and in noise, where they also assessed the relationship between age and SRTs in children aged 5 – 17 years, it was reported that the factor of age alone accounted for significant variability in SRT for children with NH (n = 63) when tested in quiet and two-talker babble (Goldsworthy and Markle, 2019).

The study by Al-Kahtani (2020) on the normal ranges of SRTs in Arabic-speaking children using the PAAST SiN looked at the mean SRTs in two age groups, 3 – 6 years (n = 20) and 6 – 12 years (n = 20). There was an age effect when comparing the two age groups (older children performed better), but there was no age effect within the younger age group (mean SRT (dB SNR) = -4.9, SD (dB) = 2.1, 95%CI (dB) = -5.9 to -3.9).

There are various factors that could lead to variations in speech and language development in children with OME, which could affect speech recognition. Number of episodes, duration and severity of OME could play a role in increasing the duration and severity of hearing loss, which could affect language acquisition (Jamal *et al.,* 2022). Other factors could be related to the home environment, where a low level of communication between parents and children in the first four years of their life could lead to decreased speech recognition (Hall *et al.*, 2014).

The current study was conducted during the COVID-19 pandemic, a few months after lockdown and the shifting of online schooling to in-person attendance to schools. Many young children spent a significant portion of their early years in lockdown, with no communication with fellow children and no regular health follow ups. A large study measuring the impact of COVID-19 pandemic on children with autism spectrum disorders, showed that lockdown had a negative impact on speech, language and behaviour in children with ASD, and as helpful as telemedicine was, many parents were having difficulties accessing these services (Bhat, 2021). These findings, along with the life course theory by Benner *et al.* (2020), support the possible impact of the pandemic on the speech, language, and mental health of young children.

The impact of enforcement of face masks as a preventive measure for all ages in closed and sometimes outdoor areas must also be considered. The mask mandate, as important as it was, may have adversely affected speech development in normal children, and in children with various types of hearing loss. Visual cues (which are decreased with masks) are considered an important element in the development of speech perception and language acquisition (Savithri, 2022). A recent study found a significant effect of both age and face masks on speech recognition in normal children aged 4-5 years, mostly due to masked acoustic signal and loss of visual cues (Kwon and Yang, 2023). The COVID-19 preventative measures may have had an effect on the speech recognition development in this group of children, possibly explaining the significant effect of age in children aged 3.5-6 years as opposed to what has been found with the study by Hall *et al.* (2007), where they found no significant difference in SRTs between NH children within the age group of 3.5-5 years.

Objective 2 To measure the ability of SRTs to predict PTA thresholds

Regarding the relationship between speech recognition and hearing levels, there was a strong positive relationship between SRTs and PTA-HL. Speech recognition in quiet represents the audibility component of hearing loss, confirmed by the results in the current study. One of the purposes of this research was to investigate the ability of the PAAST SiQ to predict PTA hearing threshold, as it is a potentially easy interactive test that can be readily applied on a computer or a smart device, and additionally, provides a real-world perception of what the child is going through in the aspect of hearing. All recruited children with OME (except for two) were able to perform PTA (n = 53), but 5 children were unable to perform the PAAST SiQ, due to factors including shyness, possibly due to restricted exposure to people other than family during COVID lock down. Another reason could be because those children did not recognise most of the pictures, possibly due to language delay which could have led to deficits in vocabulary, especially in younger children.

In the current study, an SRT of 30 dBA predicted a hearing threshold level of approximately 14 dBHL. The results from the current study were comparable with previous studies examining the relationship between speech recognition and PTA hearing threshold. Several studies assessed speech recognition using the AMTT to predict hearing levels. It was reported that an SRT of 30 dB A equated to a 14.5 dBHL (Hall *et al.*, 2007) and in another study 15 dBHL (Summerfield *et al.*, 1994).

Assessment of the relationship between average (right and left) SRTs and different PTA frequency combinations from all participants showed that all tested frequency combinations were significantly correlated with the SRTs. There was a strong correlation between SRTs and PTA-HL (n = 50), with a correlation coefficient of r = 0.7 for all tested PTA-HL frequency combination, including the 0.25 – 4 kHz combination which was not previously investigated. In both studies by Hall *et al.* (2007) and Summerfield *et al.*, (1994), all tested frequency combinations of PTA-HL of the better ear yielded a similar relationship with AMTT SRTs (obtained binaurally by loudspeakers), with correlation coefficients of approximately r = 0.7, and r = 0.8, respectively. The highest correlation with SRT was found in the study by Palmer *et al.,* (1991) with frequency combination 0.5, 1,and 4 KHz (r = 0.9). Comparisons between the current study and previous studies on the relationship between SRTs and PTA-HL of different frequency combinations are presented in Table 7.10.

Table 7.10 Comparison between the correlations between SRTs and PTA-HL frequency combinations obtained from the current study, and the studies by Summerfield *et al.* (1994), Palmer *et al.* (1991), and Hall *et al.* (2007)

PTA-HL (dBHL) Frequency Combinations (KHz)	**Current Study**		**Summerfield *et al.* (1994) Palmer *et al.*, (1991)***		**Hall *et al*. (2007)**	
	N	***r***	**N**	***r***	**N**	***r***
0.5 and 1	56	0.71	Not done		Not done	
1 and 4	56	0.77	105	0.81	962	0.7
0.5, 1 and 4	56	0.75	98	0.82	962	0.7
			66*	0.9*		
0.5, 1, 2 and 4	56	0.74	84	0.82	962	0.7
0.25, 0.5, 1, 2, and 4	56	0.71	Not done		Not done	

N= number of participants, r = correlation coefficient between SRT and PTA

Several studies looked at the relationship between PTA and speech recognition in quiet using methods other than the toy test. A study by de Andrade *et al.* (2013) found that the best correlations between PTA and SRT were found with frequency combinations 0.5 - 2 kHz and 0.5 – 4 kHz. More specifically, a study reported that in ears (n=34) with normal tympanometry results and low frequency hearing loss, frequency combination of weighted three-frequency average (0.5, 1, and 2 kHz) had the largest correlation with SRT (Kim *et al.*, 2016). Another study found that in rising audiometric configurations, the highest correlation with SRT was frequency combination 0.5 to 2 KHz (Ristovska *et al.*, 2021). These findings, along with the findings from previous studies, suggest that measuring speech recognition in general, and the PAAST SiQ specifically, can be considered as comprehensive measures of hearing that represent a wide range of frequencies, which could be utilised with PTA testing, or if PTA was not available or cannot be performed.

Objective 3 To measure the effect of OME on the AOM6

The AOM6 is an Arabic version of the OM6 questionnaire to measure QoL in children with OME. Some psychometric properties of the AOM6 have been measured in Experiment 3 (Chapter 6), including internal consistency and floor and ceiling effect. Parents of children with bilateral OME scored worse in both AOM6 (Total) and AOM6 (HaS) compared to scores submitted by parents of children with unilateral OME, but the difference between the score in unilateral and bilateral conditions was not significant. These results are similar to those from Experiment 3, where AOM6 (Total) was 3.54 (SD = 1.05) and 2.98 (SD = 0.51) in bilateral and unilateral OME, respectively. A previous study on children with OME and concurrent gastroesophageal reflux (GERD) reported that the scores of OM6 in bilateral OME were 3.5 (1.1), but their sample did not include children with unilateral OME (McCoul *et al.*, 2011).

Tao *et al.*, (2018), studied the OM6 for its ability to discriminate between the different degrees of severity of OME, factors such as number of OME episodes and antibiotic use were considered to categorise severity, but not the side (unilateral and bilateral). Another study assessed the effect of the severity of OME on QoL, defining severity by the number of OM episodes and the need for intervention. OM6 was found to be worse in recurrent OM and in children scheduled for tube placements, but the presence of bilateral OME was not considered as an indication of severity (Van Brink and Gisselsson-Solen, 2019).

Results from the current study indicated that AOM6, like the original OM6, can be considered for research as an outcome measurement for evaluating the effect of an intervention, but the severity should be evaluated by other factors such as hearing assessment and number of OME episodes. The current study provided descriptive results of AOM6 in children with OME, future studies with AOM6 data before and after treatment could provide more information on the impact of OME on QoL.

7.6.2 Aim 2 To investigate the relationship between speech recognition and QoL

Objective To measure the relationship between the AOM6 and SRTs and PTA

As mentioned in Section 2.2.7, hearing loss resulting from OME can negatively affect speech and language development and communication in children, consequently leading to adverse behavioural and psychological impacts on children, raising concerns from their caregivers. These elements are part of any evaluation of QoL in children with OME. Therefore, it is possible to find a relationship between hearing loss and QoL, mainly since the AOM6 inquires about the parents' perception of their child's hearing. A relationship between SRTs and AOM6 HaS (the average of the scores of the hearing and speech items) was also investigated, given that hearing and speech are conceptually closely connected in this age group (Heidemann *et al.*, 2013).

There was a significant correlation between SRT and AOM6 scores (r = 0.43, p = 0.002) as well as AOM6 HaS (r = 0.44, p = 0.001) scores (n = 50). Although the effect size was moderate (approximately r = 0.4 for both scores), this did not exclude the impact of hearing loss on the QoL.

The impact of hearing loss on QoL varied across studies. In a study that looked at the QoL in children with cleft palate, a group at risk for recurrent OME, hearing loss appeared to have a greater impact on QoL than physical complaints, such as earache, ear fullness, and ear discharge (De Paepe *et al*., 2019). De Paepe *et al.* (2019) attributed this to the fact that hearing loss can interfere with social communication, increasing the probability of adverse behavioural and psychological impacts, which in turn could affect school attendance and raise caregiver concerns.

A recent study on the effect of ventilation tubes in children with OME on the QoL found that all items of the OM6 were significantly decreased post operatively, the largest difference between pre-and post-operative QOL was found in the caregiver concern responses, and the least was found with speech impairment, whereas the hearing loss item exhibited the second largest difference with a significant mean decrease of 2.16 in the score (p =0.004) (Saraf *et al*., 2022).

Another aspect assessed was the relationship between PTA/SRT and the hearing item (HI) in the AOM6, to determine parents' ability to predict their child's hearing. There was a significant correlation between AOM6(HI) and average PTA (p = 0.017), but no significant correlation between the AOM6(HI) and SRTs (p = 0.114). A previous study looked at the relationship between the change in PTA-HL before and after treatment in the better ear with the change in the hearing item in the OM6, and reported no significant correlation between them (Rosenfeld *et al.,* 1998). The finding of the current study suggested that the average (right and left) PTA, considering at least one ear is affected by OME, would correlate better with the hearing item, because the information from the better ear might not represent the effect of OME on both ears. Another study found a significant relationship between the presence of CHL (nominal value, whether present or not) and hearing item of the Portuguese OM6 (p = 0.022) (Lameiras *et al.*, 2018).

Although it had been anticipated that speech recognition would have a significant correlation with the hearing item in AOM6, the fact that AOM6 (HaS) exhibited a significant correlation with speech recognition emphasises the relationship between speech recognition and its effect on the combination of hearing and speech development in children. This concept was discussed by Heidemann *et al.* (2013), where they conducted a confirmatory factor analysis (CFI) of the OM6 to investigate its one factor structure in order to test its validity in OME. The model only produced a superior fit when covariance was allowed between hearing item (Q2) and speech item (Q3), which further indicated the closeness and dependence of these items on one another. Another study that looked at the effectiveness of the parental Auditory Functional Assessment Questionnaire in predicting hearing loss in 10 children with CHL, reported no significant relationship between the parents' perception of hearing loss and their children's PTA results, further indicating that hearing loss may not be perceived correctly by parents, and that the impact of hearing loss can be noticed if parents were asked about both hearing and speech development (Fabus *et al.*, 2018).

7.6.3 Aim 3: To assess the reliability of the of the PAAST SiQ and AOM6

Objective 1 To assess the test-retest reliability of the PAAST SiQ

To determine a test is "ideal", it is required that almost identical scores are obtained when the same subject performs the test under the same conditions on two separate occasions (Lovett *et al*., 2013). To measure test-retest reliability, it is important to consider whether the differences between trials were due to the same influence (e.g., learning effect), referred to as systematic errors, or if the repeated measure gave randomly different results, namely, random errors related to subjects. The PAAST SIQ was repeated immediately (within 1-5 minutes) for the ear with the better SRT, after both ears were tested. Therefore, the factor of time was not assessed.

Systematic errors were assessed by RM-ANOVA, showing no significant difference between the two trials ($p = 0.212$), indicating good systematic variability of the test, and the results were not affected by factors such as learning effect or fatigue. Al-arfaj (2018) looked at the immediate and long-term test-retest reliability in adults using the PAAST SiQ and found a significant difference between SRTs of the first and second (immediate) measurements ($p < 0.05$) possibly due to learning effect, whereas in the long term repeat (repeat 4), there were no significant difference between the first and the fourth measurements ($p > 0.05$).

Variability of the PAAST SiQ, was assessed by within-subject variation ($SD\omega$), which was 1.95 dBA, and the repeatability (defined as the smallest difference in results when the test is repeated twice under the same condition (Lovett *et al*., 2013)) was 5.39 dB. These results were similar to what Summerfield *et al.*, (1994) found, where they reported that the AMTT in quiet in 127 children (age 2 to 13.4 years) had a variability of 2.3 dB (95%CI, 1.8 to 2.8), and a repeatability of 6.4 dB. Similar results were also reported for the AMTT in noise, where in 13 children, the variability was 3.2 dB (95%CI 1.9 to 4.6), and the repeatability was 9.0 dB (Lovett *et al.*, 2013). Al-Kahtani (2020) reported that for the PAAST SiN, the variability and repeatability were 1.6 and 4.4 dB, respectively. The small value of $SD\omega$ in the current study indicated errors that arise randomly in each subject were minimal, thus suggesting good variability and repeatability of the PAAST SiQ. Another factor that may have led to the small within-subject standard deviation is the number of reversals. In the current study, the PAAST SiQ was applied with 8 reversals, with only the last six reversal scores averaged to obtain the SRTs at 70.7%. Summerfield *et al.*, (1994) also incorporated six reversals and yielded similar results, further ensuring that the number of reversals were sufficient to decrease random errors between subjects, and at the same time is not too long for children to perform, which is advantageous in younger age groups.

The correlation coefficient can give an indication for how strongly (or poorly) correlated the results of the test are. It was anticipated that a test performed twice under similar conditions would yield a strong correlation between repeats (Lovett *et al.*, 2013). The ICC of the PAAST SiQ in the current study was 0.96, almost identical to that reported by Summerfield *et al.*, (1994) for the MTT in quiet (r = 0.95). This suggested, similar to AMTT, a strong correlation between repeats when tested under the same conditions, indicating good reliability of the PAAST SiQ.

Objective 2 To assess the test-retest reliability of the AOM6

The development and assessment of the psychometric properties of the AOM6 was achieved in Experiment 3 of the current PhD (Chapter 6), where it was shown that the AOM6 had good internal consistency, partially good construct validity and no floor and ceiling effect

An analysis of the systematic differences between the total mean scores of each repeat for all participants showed that the difference between the mean scores were close to 0, with no significant difference between scores (p = 0.433). The ICC of AOM6 in the current study was 0.7, which is the minimum standard for test-retest reliability, and grade B reliability (Gan *et al.*, 2018). Higher correlation was reported by the Heidemann *et al.* (2013) (Danish OM6) and Lameiras *et al.*, (2017) (Portuguese OM6), where the ICC was 0.85 and 0.89, respectively. Heidemann *et al.* (2013) also found that there was a systematic difference close to 0 between measurements.

An additional measure of the variations was applied to assess the random errors between subjects. The differences between measures for each subject were plotted in a Bland-Altman plot, and the results was that 93% of the data lied between ~ 2SD of the mean diffrences, only two points were outside that range, possibly due to the the narrow range between the upper and lower 2 SD, which still suggested good reliablity of the AOM6.

Although the results of the test-retest reliability of tha AOM6 were considered acceptable, one must consider stabilising the conditions between the two repeats, meaning that there might have been a change in the OME status between that first and the second repeat because of the self-resolutary nature of OME, which might have caused the individual diffrences. Based on these findings, future studies on the AOM6 with shorter intervals between test repeats may show different results in reliablity.

7.7 Summary and Conclusion

The current study aimed to ensure that speech recognition and QoL using the PAAST SIQ and AOM6, respectively, were valid reliable tools in assessing Arabic-speaking children with OME. This was investigated through several aims. One of the largest limitations in this study was that it was conducted during the COVID-19 pandemic, where restrictions, while loosened, were still in place. This limited the number of children tested and the inclusion of children free from OME. Despite these limitations, there were interesting findings, some in agreement with existing literature, and some addressing gaps in knowledge.

The main conclusions from this study were:

- During the stage of recruitment, 90% of children (n=50) were able to perform the PAAST SiQ. The remaining 10%, mostly younger children, were unable to perform the test due to shyness or inability to recognise any of the words.
- Speech recognition thresholds obtained using the PAAST SIQ exhibited approximately 5 dB significant ($p < 0.001$) elevation in ears with OME compared to ears with no OME, suggesting good discriminative validity of the PAAST SiQ in differentiating between the presence and absence of OME.
- There was no significant difference ($p = 0.572$) in SRTs between OME ears of participants with unilateral or bilateral OME.
- There was a significant effect ($p < 0.001$) of age on SRTs, which was expected based on the knowledge that children aged 3.5 to 6 vary in their normal language development. Other factors could be related to the history of OME episodes, home environment, and effect of COVID-19 lockdown.
- There was a positive significant relationship between SRTs on PTA-HL. Analysis showed a significant ($p < 0.001$) increase of 0.51 dBHL in PTA-HL with every unit (dBA) increase in SRT, suggesting that a 30 dBA SRT would equate to approximately 14 dBHL PTA-HL. All five frequency combinations of PTA-HL strongly correlated with SRTs, indicating that the PAAST can assess a wide range of hearing frequencies.
- There was no significant difference ($p = 0.601$) in AOM6 scores between children with unilateral and bilateral OME.
- There was a significant relationship between of SRTs and AOM6 (Total) ($p = 0.002$), as well as the AOM6 (HaS) ($p = 0.002$), which was a novel finding.
- There was no significant relationship between the hearing item and SRTs ($p = 0.114$), as opposed to the significant relationship between SRTs and the AOM6 (HaS), suggesting that SRTs do not represent hearing alone, but are a representation of both hearing and speech development.
- The PAAST SiQ exhibited excellent reliability, and AOM6 exhibited acceptable, moderate reliability.

In conclusion, the PAAST SiQ is considered a relatively easy tool for children aged 3.5 years and older, with good discriminative validity between presence and absence of OME in ears, and excellent test-retest reliability. Given the fact that the PAAST SiQ can to some extent predict PTA thresholds, it is possible that the PAAST SiQ, with some modifications including implementation in a smart device, can be an additional method in assessing young children's hearing. For the time being, PTA and speech recognition testing should be considered hand in hand for assessing children's hearing. Speech recognition thresholds were moderately correlated with QoL, which suggest acceptable construct validity of both tests. Future work should include testing children with no OME and comparing them with those with OME, increasing the study sample, and studying the effect of treatments on the PAAST SIQ and AOM6.

Chapter 8 Summary and conclusion

8.1 General discussion

The goal of this PhD project was to assess the use of speech recognition and QoL as outcome measures in Arabic children with otitis media with effusion. The motivation for this research stemmed from the knowledge of how common OME is among children, the nature of this condition in that it is chronic and has a high rate of natural resolution (Rosenfeld *et al.*, 2016), and its adverse effects on hearing and QoL (Homøe *et al.*, 2020). Outcome measures evaluate the effectiveness of an intervention for any disease (Chessman *et al.*, 2016), and in otitis media, outcome measures are especially important because ensuring the efficacy of a management plan can play a role in minimising short and long term complications of OME such as hearing loss, delayed language development and behavioural issues.

Recent studies have emphasized the importance of assessing hearing and QoL as part of the management of children with OME. In the current PhD, speech recognition and QoL were chosen to be assessed in Arabic children with OME. These measures were chosen for the following reasons:

- Speech recognition and QoL require language specific tools, and currently there are no validated Arabic speech recognition tests in quiet for children aged 2.5 and older, nor is there a disease specific QoL questionnaire for children with OME in Arabic.
- Given the ability of the speech recognition using AMTT to predict PTA, the PAAST SiQ (which was adapted from the AMTT) could have a potential to be used to assess hearing especially in very young children, and in facilities that lack audiometric testing.
- Speech recognition and QoL provide information on the impact of OME, both clinically and in research, and currently, information about these measures in Arabic-speaking children with OME is unavailable.

The first part of the PhD was directed towards validating the PAAST SiQ (Aims 1 and 2) and AOM6 (Aim 3) tools which measured speech recognition and QoL, respectively.

After ensuring readiness of the speech recognition and QoL assessment tools, the second part of the PhD was directed to assessing the effect of OME on these tools in order to achieve Aim 4, addressing the research question: “Can speech recognition and quality of life measured by the PAAST SiQ and AOM6, respectively, be considered good outcome measures in Arabic-speaking children with OME?”

The following sections provides a summary of each aim in the PhD project.

8.1.1 Aim 1 Equalisation of intelligibility of speech material for the PAAST SiQ test

The Paediatric Arabic Auditory Speech Test (PAAST) is an Arabic automated speech recognition test adapted from the McCormick Toy Test (McCormick, 1977), and developed by Al-Kahtani (2020). Al-Kahtani (2020) also investigated the validity of the PAAST SiN in assessing SNHL in Arabic-speaking children.

The first attempt to equalise the speech material of the PAAST SiQ was done by obtaining the parameters of the PFs of the words using MoCS in Experiment 1.A. Issues with using the MoCS, including high variability in results, warranted the use of a different method to obtain the parameters of the words of the PAAST, namely the ILAP.

The ILAP is a type of adaptive procedure that is composed of multiple tracks, each track representing a condition or a stimulus (Leek *et al*., 1991), and in the case of the PAAST, each track represented a word. Utilising the benefit of the adaptive procedures while being able to measure the SRTs of each word seemed to be the most appropriate approach towards achieving homogeneity of the PAAST SiQ, this was achieved in Experiment 1.B on NH adults. The difference from the mean of all words ranged between -3.86 and +2.15 dB, except for two words: HEN and WORMS, which had a difference of more than +5 dB. The decision was made to omit the two problematic words, and their corresponding pair words (ELEPHANT- HEN and LIGHT – WORMS). The rest of the words were equalised for intelligibility by adjusting the RMSs of the amplitude of the words by the amount of difference of each word from the mean SRT. Monte Carlo simulations (MCS) were done to assess the accuracy and precision of the PAAST when performed with different parameters. The results showed that the more reversals, the better the precision of the test, and that the Initial level should be audible, yet not too far from the expected threshold to ensure accuracy of the test.

8.1.2 Aim 2 Ensuring the PAAST SIQ is sensitive to OME related SCHL

This experiment was done to ensure that all the words of the PAAST SiQ were sensitive to CHL, because it has been suggested that eliminating any pair from the test, in case the child was not familiar with the words, would not affect its ability to detect CHL (McCormick, 1977; Summerfield *et al.*, 1994; Hall *et al.,* 2007; Lovett *et al.,* 2013). The PAAST SiQ was applied in an ILAP on 30 NH, Arabic speaking adults, who performed the test in 3 conditions: normal condition (Condition N), simulated CHL similar to that seen with OME, where low frequency thresholds were elevated (rising audiogram) and the frequencies from 2 kHz (knee point) and higher were not attenuated

(Condition 2K), and a third condition similar to Condition 2K but with a knee-point of 4kHz (Condition 4K). The ILAP allowed for obtaining the SRTs for each word in each participant, thus allowing for assessment of the effect of SCHL on each word.

The first part of the study was directed towards assessing the homogeneity of the words of the PAAST SiQ, post-equalisation. The range of SRTs was 1.95 dB, which was well below the aim range set in the pre-equalisation stage (2.5 dB), suggesting that the words were equal in intelligibility.

The second part of the study assessed the effect of the SCHL on the words of the PAAST and whether the words were equally affected by the SCHL conditions. The slopes that represented the change in SRTs from condition N to 4K were all significantly higher than 0, indicating that there was a significant effect of condition on each word. Additional analysis showed that there was a significant difference between the slopes of the words, indicating that not all words were similarly affected by conditions. The words were categorised, based on the value of their slopes, to three groups (A, B, and C), each group was composed of words with statistically similar slopes. The group with the highest slopes (A) was composed of the word pairs (BANANA – MENDRESS and BEAR-RICE). The results suggested that if any word pair needed to be eliminated, to keep at least one of the pairs from Group A to ensure the detection of slight hearing loss.

8.1.3 Aim 3 Developing an Arabic questionnaire to assess QoL in children with OME

This aim was directed towards assessing the tool for the second outcome measure of interest in the current PhD, which was QoL. Experiment 3 was conducted to develop and validate the AOM6 questionnaire. The OM6, which was developed in English by Rosenfeld *et al*. (1997), was chosen to be translated to Arabic because it was the most commonly used disease specific questionnaire for children with OME in RCTs, and it was a quick and easy questionnaire to fill (Timmerman *et al.*, 2007; Gan *et al.*, 2018). The OM6 has good psychometric properties, and its translated versions were also assessed for their psychometric properties (Brouwer *et al.*, 2005; Timmerman, 2008; Heidemann *et al.*, 2013). The OM6 was translated and cross-culturally adapted to the Arabic language following Hall *et al.* (2018) guidelines for translating QoL questionnaires. The first version of the AOM6 revealed that the words used in the question enquiring about speech delay conveyed a negative exaggerated meaning in Arabic, therefore the words were modified to ensure semantic equivalence with the original OM6 questionnaire. Parents of 42 children with an age range of 2 to 11 years diagnosed with ROM or OME filled two questionnaires: the AOM6 and the field-test questionnaires. The mean scores of the AOM6 for all participants were comparable with the scores from previous studies on the OM6 (Rosenfeld *et al*.,1997; Kubba *et* al., 2004) and translated versions (Lameiras *et al.*, 2018). The AOM6 had good internal consistency (Cronbach's alpha (AOM6 Total) = 0.72), no floor and ceiling effect, and partially good construct validity (correlation with VAS (r = -0.57, p < 0.001). The field-test revealed that most parents strongly agreed that the AOM6 was a clear, easy to understand and culturally acceptable questionnaire.

8.1.4 Aim 4 Assess the use of PAAST and Arabic OM6 in Speech recognition and QoL, respectively, in Arabic children with OME

The main aims of Experiment 4 were to (1) investigate the effect of OME on speech recognition and QoL, (2) to assess the relationship between speech recognition and QoL, and (3) to assess the test-retest reliability of the PAAST SiQ and AOM6. Arabic-speaking children aged 3.5 to 6 years with confirmed diagnosis of OME were recruited for study. Participants (n = 50) were tested by PTA and PAAST SiQ monaurally with headphones, and their parents filled the AOM6. The SRTs in OME ears were approximately 5 dB higher (significantly worse) than those in No OME ears in children with unilateral OME, but there was no difference in SRTs between OME ears of participants with unilateral and bilateral OME. There was a significant effect of age on SRTs which may have been attributed to the difference in normal development in children and the possible effects of preventative measures of COVID-19. The SRTs and PTA-HL of different frequency combinations were strongly correlated. Based on the prediction formula obtained from the parameters of the LMM, an SRT of 30 dBA is suggested to equate to approximately 14 dBHL PTA-HL average across 0.25, 0.5, 1, 2, and 4 kHz frequencies. There was a significant moderate correlation (r = 0.44, p = 0.002) between AOM6 (Total) and SRTs, indicating that as speech recognition worsens, QoL tends to worsen. The PAAST and AOM6 showed good test-retest reliability.

The findings from this study suggested that the speech recognition using the PAAST SiQ could potentially be considered as good outcome measure in Arabic children with OME, given that SRTs were increased in the presence of OME and were strongly correlated with PTA-HL. The significant correlation between the PAAST and AOM6 suggested that QoL may also be affected by OME, which could be confirmed in future studies, where AOM6 would be assessed before and after intervention.

8.1.5 PAAST SiQ and AOM6 as outcome measures in children with OME

As mentioned in Chapter 2, three main elements should be considered when assessing outcome measures.

The first is considering whether the measure would represent a change in health status. While the scope of the the current study did not cover studying the ability of the PAAST SiQ to assess the change in speech recogntion before and after OME treatment, the significant 5 dB increase in SRTs in OME ears compared to No OME ear suggests that the PAAST SiQ could potentially detect a change in hearing after treatment. The AOM6 scores were comparable to previous studies, suggesting that the scores would change after treatment of OME.

The second element is the acceptability and utility of the measures. The PAAST SiN was assessed for its usability by Al-Kahtani (2020), exhibiting a high system usability score (80/100), suggesting that the test was easy to perform. Experiment 4 (Chapter 7) showed that 90% of children approached could perform at least one PAAST SiQ run. This finding suggests that the PAAST SiQ, if implemented in a format such as a smart tablet application, could be an easy-to-use test, especially in areas where audiological equipment is unavailable. In Experiment 3 (Chapter 6), the AOM6 was found to be an easy, understandable, and culturally acceptable QoL questionnaire.

The third element is the measurement properties, which were assessed to some extent throughout the current thesis for both measures. The PAAST SiQ was assessed for its homogeneity and sensitivity to simulated hearing loss before applying it to children. The PAAST SIQ showed good discriminative validity between OME and No OME ears and good test re-test reliability. The AOM6 was also assessed for its psychometric properties, where it showed good reliability, no floor and ceiling effects, and partially good construct validity.

The PAAST SiQ and AOM6 can potentially be considered good, reliable outcome measure tools to be further researched for assessing children with OME.

8.2 Conclusions

The main conclusions of this PhD project were:

- Ensuring equal intelligibility of speech material is an important step in the development of speech tests and failing to do so may affect the accuracy of the results. Also, the homogeneity of the speech material can be affected by the presence or absence of noise background. Therefore**,** it is important to equalise the intelligibility of words in both noise and quiet.
- The method of obtaining SRTs in a speech test can affect the accuracy of the results. Therefore, it would be worthwhile to perform a pilot study to assess the efficacy of the method used. In the current study, the interleaved adaptive procedure (ILAP) resulted in more accurate and less variable results in the PAAST SiQ compared to the MoCS.
- This research addressed a gap in knowledge regarding the sensitivity of each word in the PAAST SiQ to OME-related SCHL. The results showed that all the words in the PAAST SiQ

were sensitive to SCHL, with some words showing more sensitivity than others. This suggested that elimination of any word in the PAAST SiQ, would possibly not affect the sensitivity of PAAST SiQ in detecting OME related CHL.

- This research was the first to address QoL in children with OME in the Arab world in general, and in Saudi Arabia specifically. The AOM6 was adapted from the validated, disease specific OM6 questionnaire and was assessed for its validity. The AOM6 has shown to be a reliable, valid tool to assess children with OME, but further validation in the future is required by assessing additional psychometric properties including responsiveness to intervention.
- The PAAST SiQ and AOM6 both showed good test-retest reliability.
- Speech recognition and QoL were assessed in children with OME, addressing yet another gap in knowledge in the Arab world, providing data of both measures in Arabic-speaking children with OME. The PAAST SiQ was able to distinguish between ears with OME and ears with no OME, with a significant 5 dBA increase in OME ears, but there was no significant difference in SRTs in OME ears in children with unilateral OME compared to those with bilateral OME.
- The PAAST is an easy, reliable test that could be implemented as an application on smart devices (tablets). Also, given that PTA thresholds could be predicted from SRTs suggests that the PAAST SiQ could potentially be used in testing children as young as 3 years old who can't perform PTA, or in clinics where PTA is not available.
- There was a significant effect of age on SRTs in the group of children aged 3.5 to 6 years, with a 6 dBA improvement per year. This effect could be due to factors related to natural development of vocabulary or history of the OME disease (which were not investigated in the current study) but could also be attributed to the effect of COVID-19 lockdown, which most children in the study experienced at an early age.
- The relationship between QoL and PTA thresholds was assessed in previous studies, but the current PhD explored the relationship between SRTs and AOM6 total scores. A significant relationship was found between the two measures, but there was no significant relationship between the hearing item of the AOM6 and speech recognition. This suggests that a comprehensive approach to investigating hearing in a child with OME should include hearing tests (including PTA and speech tests) and QoL assessment, without complete reliance on the parents' perspective of their child's hearing.
- The PAAST SiQ and AOM6 could potentially be considered as good outcome measure tools to assess speech recognition and QoL, respectively. Ideally, further validation studies should be conducted before using these tools in clinics or in RCTs that investigate the effects of treatment of OME on the speech recognition and QoL in Arabic children.

8.3 Limitations and Covid Statement

Most limitations to the current PhD project were attributed to the COVID-19 pandemic where the following issues were faced:

- The lockdown mandates were implemented during the data collection phase of Experiment 1.B, leading to termination of the data collection after n = 17 participants were tested, as opposed to the originally set sample size of n = 20.
- The data collection of the AOM6 validation study (Experiment 3) was also paused after 5 children had been tested, due to the lockdown. The experiment was revisited after lockdown mandates were lifted, but the recruitment process was affected by the pandemic.
- The uncertainty of the COVID-19 situation, lead to considering different approaches to address the aim of the PhD, including experiments that did not require data obtained through direct contact with humans. The time from March to August 2020 was spent writing up, performing computer simulations (Experiment 1.B) and revisiting the aims of the PhD.
- Originally, Experiment 1.B included a post-equalisation study to assess the homogeneity of the PAAST SiQ. This step was performed as part of Experiment 2, to allow time to continue with the remainder of the studies.
- The decision was made to revisit Aim 4, which was initially to be addressed by conducting a case-control study of children with OME and age and sex-matched NH children. The revisited final study design was to assess speech recognition and QoL in children with OME who were coming for their appointments in the ENT clinic. Normal children were excluded to avoid having them come to a hospital during the pandemic without a medical reason.

8.4 Future work

The recommended future works is as follows:

- The current PhD study provided preliminary data in children with OME when tested with the PAAST SiQ and AOM6. However, conducting a similar study with larger sample size with different age groups and demographics, as well as including history of the disease, for example, the number of OME episode in the past, can provide information about how environmental and disease factors can influence speech recognition and QoL.

- A study that compares children with OME with age and sex – matched NH children could provide information on how speech recognition differs between groups. It is recommended to assess NH children of different age groups with the PAAST SiQ to obtain normal data on speech recognition.
- Assessment of speech recognition in OME children in a free field using speakers at different positions could provide information on the effect of OME/prolonged OME on localisation.
- Assessment of OME children with both PAAST SiQ and SiN could provide information on the impact of OME on listening in different conditions, bearing in mind inclusion of disease history and environmental factors of the child.
- The work on the validation of the AOM6 questionnaire is recommended to be continued. Conducting a study with a large sample size ($n \geq 100$) to perform PCA and measure psychometric properties such as discriminative validity and responsiveness by assessing children before and after surgery would allow for ensuring the validity of the AOM6 and obtain cut-off scores.
- Studying the pattern of responses of the PAAST could provide information about whether participants would wrongly choose a word in a certain pair (at low intensities) when the presented word was the other word from the same pair. Studying these patterns could provide information as to whether the PAAST is truly a discrimination test.
- The ability of the PAAST SiQ and AOM6 to detect change in speech recognition and QoL, respectively, is recommended to be assessed in a longitudinal study that applies these tests before and after treatment of OME, whether the treatment was conservative or interventional (e.g., surgery). Ensuring these measures are sensitive to change in OME status could allow for using the PAAST SiQ and AOM6 as outcome measure tools in RCTs.

Appendix A Audiological testing in children

A.1 Types of Audiological tests in children

Table A.2 Audiological testing for infants and children adapted from (Harlor *et al.*, 2009; British Society of Audiology, 2018; Farinetti *et al.*, 2018)

Developmental Age of Child	Auditory Test	Measurement Type	Advantage	Limitations
All ages	Evoked OAEs	**Objective** measure of the function of outer hair cells in response to a stimulus	Ear specific, quick, easy, does not require patient to be sleeping Screening test	Child must be relatively inactive, and it is very sensitive to OME and cerumen, and could be normal in cases of auditory neuropathy
All Ages	Auditory Brainstem Response (ABR)	**Objective** Electrophysiological measurement of neural activity in the auditory nerve and the brainstem	Ear specific, not dependent on patient cooperation. ABR is diagnostic. Automated ABR is for screening	Child must be quiet, sedation is often required
Birth to 9 months	Automated ABR			
6 to 36 months	Distraction test	**Subjective** behavioural test that relies on detracting the child while presenting a stimulus using a sound emitting object (e.g. rattle) or a hand held sound generator	It allows assessing hearing sensitivity in infants who are unable to perform behavioural testing reliably using VRA.	Hearing threshold may not be accurate
9 months to 2.5 years	Visual Reinforce-ment Audiometry (VRA)	**Subjective** behavioural test measuring response of child to frequency specific stimulus presented through speakers or insert-phone and the child is rewarded with a visual reinforcement	Assess auditory perception of the child Diagnostic test	When performed using free-field speakers it can only assess the hearing in the better ear
2.5 to 4 years	Play audiometry	**Subjective** behavioural test measuring auditory thresholds in response to frequency specific stimulus presented through earphones phones or bone vibrator.	Assess auditory perception of child; screening or diagnostic	Attention span of the child may limit amount of information obtained

Appendix A

Developmental Age of Child	**Auditory Test**	**Measurement Type**	**Advantage**	**Limitations**
4 years and older	Conventional audiometry	**Subjective** behavioural test measuring auditory thresholds in response to frequency specific stimulus presented through earphones phones or bone vibrator.	Assess auditory perception of child; screening or diagnostic. It is considered a gold standard for testing hearing (Farinetti *et al.*, 2018)	Depends on the level of understanding of the child

Adapted from (Harlor *et al.*, 2009)

Appendix B Studies on the QoL in children with OME

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
Surgery for Otitis Media in a Universal Health Care Model: Socioeconomic Status and Race/Ethnicity Effects (Ambrosio and Brigger, 2014)	San Diego, USA English	Prospective	(1) To determine the association between socioeconomic status (SES), race/ethnicity, and other demographic risk factors in surgically managed otitis media within a model of universal health care. (2) To determine QoL outcomes of surgically managed otitis media in this model.	1ry: to determine the role of SES, race/ethnicity and other demographic risk factors in surgically managed OM 2ry: measure QoL outcome	240 (120 cases) (120 controls)	<6 years	OM-6	***p-value (control vs Post-op score)*** PS .931 , HL .509, SI .860, ED .340 , AL .297, CC .807 Overall .490 -no difference between control and post op group -significant change in all domains pre- and post op
Quality of Life after Surgery for Recurrent Otitis Media in a Randomized Controlled Trial (Kujala *et al.*, 2014)	Oulu, Finland Finnish	RCT	**Method:** apply the questionnaires to 3 randomized groups, tympanostomy, tympanostomy with adenoidectomy and no surgery; at diagnosis, 4m and 12m	1ry outcome: to evaluate the QoL of children with RAOM after tympanostomy	n=159	10m – 2y	OM-6 Visual scale of ear-related QoL	Significant improvement in the 12m F/U in the subset of OM-6: caregiver concern, emotional distress and physical suffering Tympanostomy did not have affect on QoL @ 12m compared to no surgery **Conclusion: the** QOL of children with RAOM may improve with time – no difference with surgery

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
Impact of Tympanostomy Tubes on Child Quality of Life (Rosenfeld *et al.*, 2000)	USA English	Prospective	**Objective:** To determine the subjective impact of tympanostomy tubes on child QOL, and to compare the variability in QOL before surgery with that observed after surgery. **Method:** OM-6, Satisfaction with Decision Scale, and satisfaction with office visit; surveys were completed at baseline (visit 1), at surgery (visit 2), and after surgery (visit 3).	Short-term changes in QoL.	248	6m-12y **median** 1.4y	OM-6	Large, moderate, and small improvements in QOL occurred after surgery in 56%, 15%, and 8% of children, respectively.
Adenoidectomy plus tympanostomy tube insertion versus adenoidectomy plus myringotomy in children with obstructive sleep apnoea syndrome (Vlastos *et al.*, 2011)	Athens, Greece Greek	RCT	**Objective:** To determine whether T- tube insertion has benefit, compared with simple myringotomy, in children with OME **Method:** QoL questionnaire was applied to randomly assigned children in two groups: (1) adenoidectomy +tympanostomy tube (2) adenoidectomy + myringotomy	- QoL 6 and 12 months post-operatively -audiometric threshold	52	>3 years	OM-6	OM-6: group (1): significant change in OM-6 score ($p < 0.05$) Six months after surgery compared to pre-operative. No significant change at 12 month point.

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
Randomised controlled trial of the effect of ventilation tubes (grommets) on quality of life at age 1-2 years (Rovers *et al.*, 2001)	English	RCT	**Aims:** To study the effect of treatment with VT on QoL in children aged 1–2 years with persistent OME, as compared to watchful waiting **Objective:** to apply QoL questionnaires at 0, 6, and 12 months follow up	-quality of life -hearing assessment	187	9-12 months	TAIQOL and Erickson scales	
Grommets (ventilation tubes) for recurrent acute otitis media in children. (Venekamp et al., 2018)	English	Meta-analysis	**Objective:** To assess the benefits and harms of bilateral grommet insertion with or without concurrent adenoidectomy in children with rAOM **Method**: search databases	**1ry:** AOM recurrence (intermediate term) **2ry:** - AOM recurrences (long-term) -total number of AOM recurrences -quality of life			OM-6	***Grommet vs. active monitoring*** 1 RCT showed that children receiving grommet tubes did not have better QoL (OM-6) at 4 or 12 months than those managed by active monitoring ** low quality due to it being 1 study with low number of participants (85)

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
Interventions for children with ear discharge occurring at least two weeks following grommet (ventilation tube) insertion (Review) (Venekamp *et al.*, 2016)	English	Meta-analysis	Objective: To assess the benefits and harms of current treatment strategies for children with ear discharge occurring at least two weeks following grommet (ventilation tube) insertion	**1ry:** - resolution (short term) -adverse outcome **2ry:** - resolution (long term) -quality of life -hearing			**Generic:** Child health questionnaire (CHQ) **Disease Specific:** OM-6	***Quality of life in children treated with antibiotics vs, those treated with antibiotic drops:*** Difference in change in median OM-6 scores: -2 (in favour of antibiotic eardrops)
A Trial of Treatment for Acute Otorrhea in Children with Tympanostomy Tubes (Van Dongen *et al.*, 2014)	Nether-lands Dutch	Open-label RCT	Comparison of the effectiveness of three strategies for the management of acute tympanostomy-tube otorrhea in children: (1) immediate treatment with antibiotic–glucocorticoid eardrops (76), (2)immediate treatment with oral antibiotics (77), and (3)initial observation (77).	**1ry:** presence of otorrhea after 2 weeks **2ry:** -duration of otorrhea -the total number of days of otorrhea -number of otorrhea recurrences after 6 months -quality of life -complications, - adverse events	230	1-10 years	**Generic:** Child health questionnaire (CHQ) **Disease Specific:** OM-6	***Quality of life in children treated with antibiotics vs, those treated with antibiotic drops:*** Difference in change in median OM-6 scores: -2 (in favour of antibiotic eardrops)

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
Quality of life outcomes after ventilating tube insertion for otitis media in an Australian population (Chow, Wabnitz and Ling, 2007)	English	Prospective pre- and post-intervention outcome study	**Objective:** To assess the change in QoL in a group of children with recurrent (AOM) and or OME treated with ventilation tube (VT) insertion **Method:** OM-6 was applied prior to surgery and 6 weeks after surgery	Quality of life	53	11 months to 15.4 years	OM-6	***Quality of life improvement after surgery:*** significant improvement in the mean score of OM-6 score (p*<0.001) and all domains of the OM-6*
Oral steroids for hearing loss associated with otitis media with effusion in children aged 2–8 years: the OSTRICH RCT (Francis *et al.*, 2018)	English	Double-blind, individually randomised, placebo-controlled trial	**Objectives:** to determine the effects of 7 days steroid therapy on HL associated with OME **Method:** A 7-day course of oral soluble prednisolone, as a single daily dose of 20 or 30 mg for 6- to 8-year-olds, or matched placebo	1ry: improved hearing at 5 weeks 2ry: -improved hearing (long-term) -clinical exam -proportion of children receiving VT -quality-of-life	389	2–8 years	PedsQL OM8-30	No signican differeance in the PedsQL scores between the groups

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
The prevalence and impact of otitis media with effusion in children admitted for adeno-tonsillectomy at Dr George Mukhari Academic Hospital, Pretoria, South Africa (Els and Olwoch, 2018a)	English to non-English speaking populatio n (south-african) in assistance of the researcher and Setswana speaking nurse.	A cross-sectional, observa-tional study	**Objective:** - prevalence of OME in children admitted for adeno-tonsillectomy - impact of OME on QoL	Hearing loss Quality of life	109	2–12 years mean = 6.1	OM-6	The mean total OM-6 survey score was 1.67 (SD ± 0.59) in children with OME, and 1.31 (SD ± 0.45) without OME, showing no statistically significant difference ($p > 0.05$)
Evaluation of Children Quality of Life after Serous Otitis Media Surgery (Jabbari Moghaddam and Mirghaffari, 2018)	Iran, Farsi	prospective cross sectional study	**Aim:** To examine the changes in the children's QoL after OME surgery **Objectives:** to measure the results of the OM-22 filled by parents before and 12-weeks after surgery.	Quality of life	70	1 to 12 years	OM-22 (translation to Farsi, content validity and reliability done)	Significant decrease in the OM-22 scores after surgery; i.e. improved QoL.

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
The effect of ventilation tube insertion to the health-related quality of life in a group of children in Southeast Anatolia (Yazici and Coskun, 2018)	Turkey, Turkish	Prospective	**Aim:** To demonstrate the influence of ventilation tube insertion to the QoL in children with OME **Objective:** to apply OM-6 1 week before and 6 weeks after surgery	Quality of life	45	Mean = 67.64 ± 42.89	OM-6 (forward and backward translation)	Moreover, each domain of OM6 showed statistically significant improvement before and after surgery
Quality of life in Swedish children receiving grommets – An analysis of pre- and postoperative results based on a national quality register (Van Brink and Gisselsson-Solen, 2019)	Sweden, Swedish	Cohort	**Aim:** to address the effect of surgery in children with OME on the QoL. Objectives: to analyse data of QoL questions answered by parents pre- and post-ventilation tube insertion	Quality of life	307	Mean = 4.4 years	QoL questions with the national Swedish quality registrar. No specific QoL questionnaire used.	Significant improvement of QoL compared between pre and post-surgery

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
Middle ear disease in Danish toddlers attending nursery day-care – Applicability of OM-6, disease specific quality of life and predictors for middle ear symptoms (Indius *et al.*, 2018)	Denmark, Danish	Cross-sectional	**Aim:** to investigate the difference in QoL between three groups of children; Children with symptoms of ear disease within the last 4 weeks, children without any ear disease and children scheduled for ventilating tube treatment **Objectives:** All groups filled the Danish OM-6 questionnaire and results were analysed.	Quality of life	342	21.1 months	Danish OM-6 (this was a validation study as well)	QoL was significantly worse in the 4-week group compared to the non-4week group.
Effect of Pneumococcal Vaccination on Quality of Life in Children With Recurrent Acute Otitis Media: A Randomized, Controlled Trial (Carole N.M. Brouwer, Maillé, Rovers, Veenhoven, *et al.*, 2005)	Nether-lands Dutch	Double-blinded RCT	**Aim:** To assess the effect of Pneumococcal vaccination on HR-QoL or FHS **Objectives:** children with RAOM were vaccinated with either heptavalent pneumococcal conjugate vaccine followed by pneumococcal polysaccharide vaccine (pneumococcal group: n=190) or with hepatitis A or B vaccines (control group: n = 193)	Quality of life	383	1-7 years	***Generic FHS*** ***OM–specific FHS:*** OM6 ***OM–specific child HR-QoL*** -Numerical Rating Scale for Child - Family Functioning Questionnaire -Numerical Rating Scale for Caregiver	AOM frequency decreased 4.4 episodes per year in both groups, with a considerable and comparable improvement in HR-QoL and FHS

Study Title	Language	Study Type	Objective and method	Outcome	N	Age	Questionnaire	Result
Impact of Otitis Media Severity on Children's Quality of Life (Blank *et al.*, 2014)	USA, English	cross-sectional study	**Aim:** quantitate the average burden of OM and to compare the associated impact of tympanostomy tubes on infant health related quality of life (HR-QoL) **Objectives**: case group: children with recurrent AOM or COM Control group: children who are otologically normal PedsQL and OM-6 was applied to both groups	HR-QoL	1208	6 to 24 months Mean = 14.7 months	PedsQL OM-6	-mean OM-6 score of children with recurrent OM was 3.3, whereas similarly aged well-children had a mean OM-6 score of 2.5 - Worse OM-6 scores were correlated with poorer PedsQL Infant scores, Pearson *r* = –0.581 (1-12 months) and –0.558 (13-24 months), P < .001
Quality-of-Life Outcomes After Surgical Intervention for Otitis Media (Richards and Giannoni, 2002)	USA, English	Prospective	**Aim:** To assess the change in QoL in children with rAoM and/or COM treated with surgical intervention **Objectives:** QoL questionnaire was administered before and after surgical intervention	HR-QoL	123	<16 years	OM-22	The mean percentage change after surgery: -total ear symptoms score: 74.5% improvement (P<.001) - Parental worry "mean score" 3.43 (P<.001)

Appendix C Participants' criteria for Pilot study and Experiment 1.A, 1.B and 2

C.1 Inclusion and exclusion criteria for participants

Inclusion criteria	Native Arabic speaking adults
	18 years – 45 years
	PTA air-conduction threshold at frequencies (0.125 - 8 kHz) $\leq$ 15 dB HL
	Ear examination with otoscopy: normal external and middle ear
	Tympanometry: Type A
Exclusion criteria	Age < 18 years and > 45 years
	History of sensorineural hearing loss
	History of recent ear infections or ear surgeries
	External ear canal abnormalities (e.g. impacted wax)
	Middle ear abnormalities (e.g. otitis media with effusion)
	Any type other than Type A in Tympanometry

C.2 Otological health questionnaire

Institute of Sound
and Vibration Research

UNIVERSITY OF Southampton

Otological Health Questionnaire

Study Title: Testing the Sensitivity of the Arabic to Test in Quiet to Simulated Conductive Hearing Loss Caused by Otitis Media with Effusion

Researcher: Sarah Alsebai **Ethics number:** 46958

ID of participant ……………………………………….

Age ………………… **Gender** ……………………………….

Date …………………………

	Please circle	
Have you had any ear disease (infection, persistent pain, discharge) in the last 6 months?	Yes No	If Yes please provide further details ______________________ ______________________
Have you ever had any ear surgery?	Yes No	If Yes please provide further details ______________________ ______________________
Are you aware of any family history of hearing loss?	Yes No	If Yes please provide further details ______________________ ______________________
Do you suffer from tinnitus? (ringing in the ear)	Yes No	If Yes please provide further details ______________________ ______________________
Have you been exposed to any loud noise over the past 48 hours?	Yes No	If Yes please provide further details ______________________ ______________________
Do you have any concerns about your hearing?	Yes No	If Yes please provide further details ______________________ ______________________

Appendix D Speech Recognition and the psychometric functions

D.1 Speech test Audiometry: Methods of testing and the psychometric functions

Several parameters are used to describe the results of speech audiometry. One of the main parameters that is commonly used to assess speech recognition is the speech recognition threshold (SRT), which is defined as the level of intensity at which individuals can repeat words 50% of the time (Carhart, 1951), and is reported in the unit by the same unit as the stimulus intensity; either dB A in quiet or dB SNR in noise. Other parameters can be assessed from a speech audiometry such as speech recognition score; which represents the percent correct recognition at suprathreshold levels, and the roll over-index; which represent the index of roll-over of the performance, this can be used to diagnose retro-cochlear causes of hearing loss (Katz *et al.*, 2009). The speech test audiometry parameters can be represented as psychometric function which is performance i.e. % correct plotted as a function of intensity (dB A) (Figure AD.1).

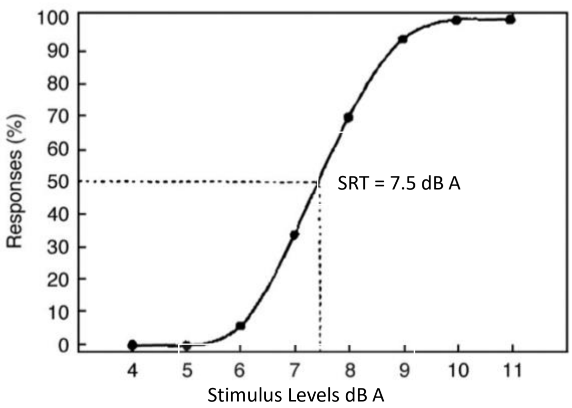


Figure AD.2 Psychometric function obtained using the constant stimuli method (adapted and edited from Gelfand, 2009).

In order to construct the psychometric function for an individual, a method for obtaining the data needs to be chosen, this will be described in the next section.

D.2 Method of testing

Several methods are available to obtain data to be later fit to a psychometric function (PF), including the method of constant stimuli (MoCS), method of limits and adaptive procedure (Gescheider, 1997). The Method of constant stimuli is commonly used in validating and testing the sensitivity of a speech test (Summerfield *et al.*, 1994; Semeraro *et al.*, 2017), and is defined as "the procedure of repeatedly using the same set of stimuli (usually between five and nine different values in the set) throughout the experiment". The levels are selected based on previous experiments or pilots. The range of levels is chosen so that it encompasses all possible intensities which would result in responses from 0% to 100%. The speech material lists are presented numerous times at each level in a random order. The results are then plotted and fitted into a psychometric function, details of the elements of the PF functions are in explained in the next section.

D.3 The psychometric function

To visualise the relationship between the acoustic stimulus and response, a psychometric function, which is an "S" shaped tracing, represents the function of the intensity level (X-axis) and the percentage of correct responses (Y-axis). The "S" shape, sometimes referred to as ogive (Gescheider, 1997), is a representation of what one would experience when listening to very faint sounds around the threshold level, where the response could be as low as 0% (lower asymptote), and the path it takes when the intensity level increases, where the response level improves until it reaches 100% (upper asymptote) (Carhart, 1951).

The MoCS was chosen to obtain the data, that were later fit and analysed in the psychometric function following these steps (Kingdom and Prins, 2016):

1- **Choosing the stimulus levels**

   Typically, the set of stimulus levels chosen must result in a response that ranges from just above chance to nearly 100%.

2- **Selecting the function to fit the data**

   Several functions are available including logistic regression, cumulative normal, Weibull and hyperbolic secant.

   For this experiment in this report, the logistic function was chosen by convenience to fit the data points.

3- **Fitting the function**

   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).

The parameters estimated from the fitted function are shown in Figure AD.2.

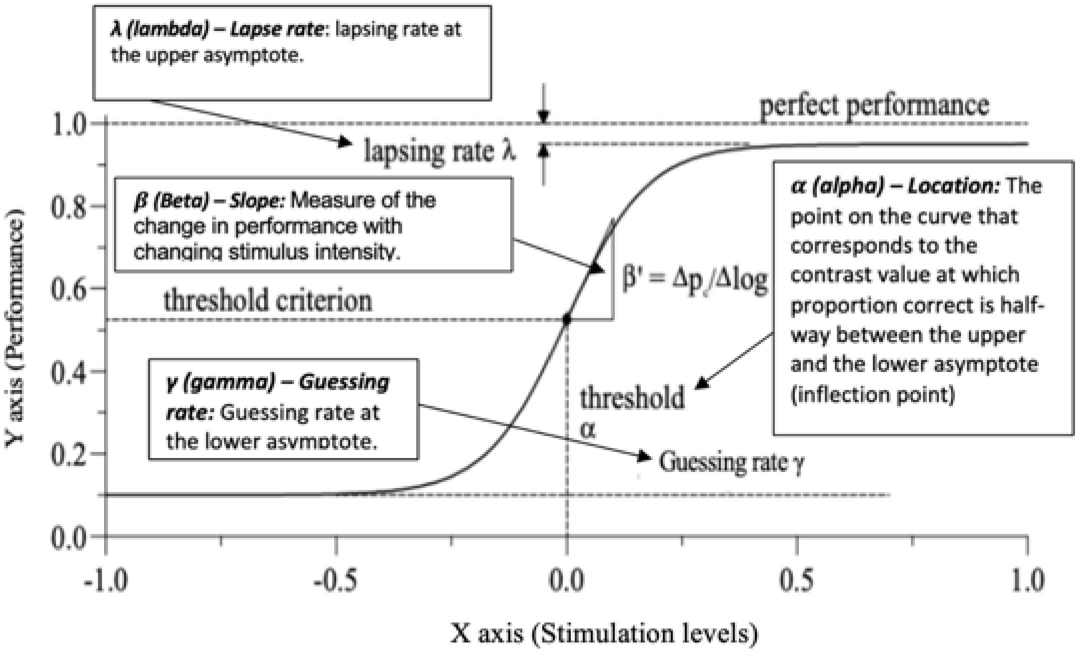


Figure AD.2 Psychometric function and its components (adapted and modified from Strasburger, 2001).

- α - Location: is the position along the abscissa and is defined as a certain value, which corresponds to the inflection point which we will term “the location”. The location is the steepest point in the slope and is reported in dBA. Several researchers have chosen the reference point to be SRT50 (SRT that corresponds to 50% correct) such as Semeraro et al. (2017), Ozimek et al. (2010) and Smits, Kapteyn and Houtgast (2004). The reason why location was chosen as a reference point in the current PhD rather than SRT50 is that the latter may not always represent the inflection point (steepest point), as opposed to the location which does, which provides a generalised approach to assessing the parameters of the test/speech material for different purposes. without limiting the parameter to a certain percent correct.
- β corresponds to the slope. The slope expresses the ratio of maximum change in performance to the change in stimulus level intensity. It is a measure of the function’s precision (Kingdom and Prins, 2010). The slope is the gradient of the PF, and a single value can’t describe the slope because is different across the PF. The value of β is a

theoretical concept that is equivalent to the slope of a tangent line at the point of infliction of the PF (Semeraro, 2015). Typically, it is calculated from the derivatives of a secant line between the point of inflection and a "close point" referred to as (*s*), divided by the distance between them on the x-axis (*h*) (Figure AD.3). The closer the *s* is to the inflection point and the closer the *h* to 0, the closer the secant line to the tangent line, and the better approximation of the slope (Semeraro, 2015).

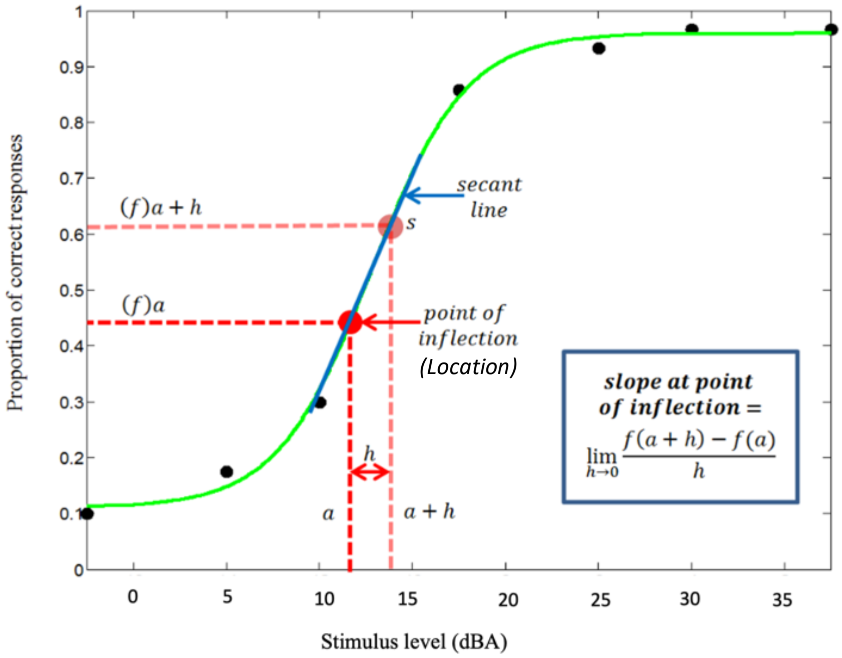


Figure AD.3 Obtaining the slope of the infliction point of a psychometric function (adapted from Semeraro (2015))

Another way to express the slope of PF for speech recognition is by finding the relation between the change in correct recognition performance (Ay) and the change in the presentation level of the signal (Ax) that is expressed as some form of Ay/Ax (%/dB) (Wilson and Carter, 2001). This translates to a steep slope indicating large changes in intelligibility occurring due to small changes in SNR while shallow slopes indicate that larger changes in stimulus level are required to elicit the same change in intelligibility as the steep slopes (Macpherson and Akeroyd, 2014).

- 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 (0.01) would prevent significant effect of bias on the location and slope.

4- Estimating the errors on the function's parameter estimates

Standard errors (SE) can be calculated through a process called bootstrap analysis. Two things can influence SE:

a. Sample size
b. Amount of variation

The smaller the SE, the smaller the distance of the data from the psychometric function, which means acceptable results

5- Determining the goodness-of-fit of the function

Goodness-of-fit is a measure of how well the data fit the psychometric function. This measure is a p-value (pDev) that will always have a value between 0 and 1; the greater the value, the better the model describes the data.

6- Confidence intervals

Standard errors (SE) and their confidence intervals (CI) must be considered alongside the pDev for accurate interpretation of the results. Neyman (1937) was the first to introduce the concept of confidence intervals into the science of statistics. The follow definition is adapted from the original definition with modification relevant to the topic being discussed. The 95% confidence interval (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. The 95%CI was calculated as:

$$95\%CI = 2 \times SE$$

Appendix E Experiment 1.A: Additional data

E.1 Method of constant stimuli Levels in dBA for each condition for pilot

	N1 Normal Condition	R1 SCHL with cut-off point 2 kHz	R2 SCHL with cut-off point 4 kHz	R3 SCHL with cut-off point 8 kHz
Levels in dBA	16	22	26	29
	12	18	22	25
	10	16	20	23
	8	14	18	21
	6	12	16	19
	2	8	12	15

E.2 Parameters of the PFs for each word across participants in the Pilot study

Words	Location (dBA)	Difference of location from the mean location (dBA)	SE of Location (dBA)	Slope	Slope in %/dB	pDev
BANANA	10.8	1	1.2	0.5	11.5	1
DOG	6.7	-3.3	2.1	0.3	8.3	0.98
DOOR	2.8	-7.1	7.9	0.3	7.8	1
ELEPHANT	10.5	0.6	1.4	0.3	8.6	0.99
EYE	7.6	-2.3	2.2	0.4	9.5	1
BEAR	8.3	-1.6	1.3	0.3	8.7	1
HEN	16	6.1	592.1	0.2	4.6	0.99
HOUSE	12	2.1	91.1	0.3	8.5	1
LIGHT	10.4	0.5	1.8	0.3	7.4	1
MENDRESS	7.7	-2.2	1.4	0.4	10.4	0.99
RICE	13.5	3.6	157.8	0.3	8.2	1
PEOPLE	11.8	1.9	315.2	0.3	6.7	1
FLOWER	7.5	-2.4	1.5	0.3	8.7	1
WORMS	13	3.1	447.4	0.3	6.4	1
Mean	9.9			0.32	8.24	1.00

(Highlighted in grey are words with SE ≥8 dB which was unacceptable)

E.3 Studies that equalised the intelligibility of speech materials in noise and in quiet

Study	Test	Language	Number of Participants	Findings
Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise (Nilsson *et al.*, 1994)	HINT	English	6-8 subjects per round, 7 rounds	Method: to present the sentences in noise (using the spectrally matched noise) at a fixed SiN ratio to normal-hearing listeners and to measure percent intelligibility 1 dB per 10% difference Equalisation done in noise (but not in quiet) *SRTs in quiet are correlated highly with the pure-tone thresholds
Clinical evaluation and test-retest reliability of the IHR-McCormick Automated Toy Discrimination Test (Summerfield *et al.*, 1994)	McCormick Toy Test	English	8 participants	The levels of words producing 71% correct responses ranged between -2.5 and +2.9 dB SNR and -3.9 and +5.0 dB in quiet. Equalisation was done in noise and quiet separately
Development of the Cantonese Hearing In Noise Test (CHINT) (Wong and Soli, 2005b)	CHINT	Cantonese	6 subjects per round, 5 rounds (30 participants)	The overall RMS level of each sentence was changed by 1 dB for each 9.7% below or above the 65% reference point

Study	Test	Language	Number of Participants	Findings
Inter-list Equivalency of the CID W-22 Word Lists Presented in Quiet and in Noise (Loven and Hawkins, 1983)	CID W-22 Word Lists	English	48 subjects (NH)	(1) the lists were equivalent when administered in quiet; (2) the lists were not equivalent when administered in a background of multi-talker babble; (3) the addition of noise changed the relationship among the word lists in a unpredictive manner; (4) the addition of noise changed the level of difficulty of some words of each list to a greater degree, relative to the other words
Inter-list Equivalency of the Northwestern University Auditory Test No. 6 in Quiet and Noise with Adult Hearing-Impaired Individuals (Stockley and Green, 2000)	NU-6	English		The levels required to equalise the intelligibility of the words in quiet are different from those used to equalise the intelligibility in noise.
Variations in the Slope of the Psychometric Functions for Speech Intelligibility: A Systematic Survey (Macpherson and Akeroyd, 2014)	The threshold and the slope of the psychometric function for speech intelligibility can vary with different conditions, e.g., quiet and noise, as well as different maskers in noise. Single speech maskers are likely to give particularly shallow slopes.			

E.4 The levels used for the MoCS in the PAAST in quiet and the PAAST in noise

dBA levels in quiet	dB SNR levels in noise
20	-7
18	-9
16	-11
14	-13
12	-15
10	-17
8	
6	
4	

E.5 Results of the psychometric function of the PAAST in quiet – Experiment 1.A

	Location (dBA)	SE (dB)	95%CI upper limit (dB)	95%CI Lower limit (dB)	Difference from mean location (dB)	pDev	Slope	Slope (%/dB)
BANANA	15.04	1.07	16.95	-12.69	2.08	1	0.4	10.95
DOG	11.48	1.11	13.26	-9.18	-1.48	1	0.41	11.21
DOOR	9.03	1.38	11.82	-6.21	-3.93	1	0.32	8.18
ELEPHANT	12.92	1.09	14.75	-10.86	-0.04	1	0.41	11.48
EYE	8.91	1.05	10.9	-6.79	-4.05	1	0.44	11.73
BEAR	9.83	1.1	11.45	-7.22	-3.13	1	0.38	10.4
HEN	19.47	ERROR	ERROR	ERROR	6.51	1	0.19	4.6
HOUSE	11.24	1.13	12.96	-8.76	-1.72	1	0.38	10.5

	Location (dBA)	SE (dB)	95%CI upper limit (dB)	95%CI Lower limit (dB)	Difference from mean location (dB)	pDev	Slope	Slope (%/dB)
LIGHT	15.54	1.46	17.72	-12.6	2.58	1	0.3	7.99
MENDRESS	12.63	1.06	14.08	-10.21	-0.33	1	0.41	12.09
RICE	16.58	1.8	19.15	-13.64	3.62	1	0.3	8.04
PEOPLE	11.28	1.35	14.17	-8.12	-1.68	1	0.28	6.76
FLOWER	9.39	1.08	11.45	-7.16	-3.57	1	0.42	10.23
WORMS	18.09	ERROR	ERROR	ERROR	5.13	1	0.24	6.29
Mean	12.75						0.35	9.32

locations in dBA for each word, difference from the mean location in dB, SE: Standard Error, 95 CI: Confidence interval, slope, %/dB: change of function per dB, pDev: goodness-of-fit

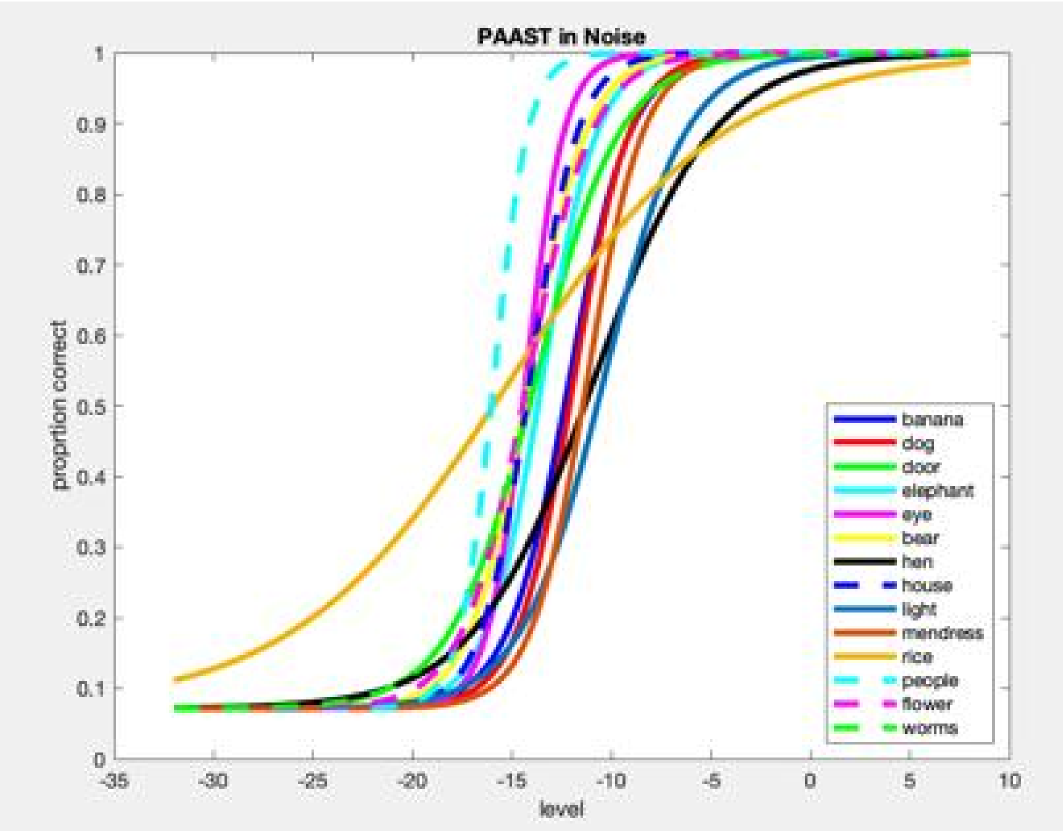
E.6 Psychometric functions results of the PAAST SiN from Experiment 1.A

E.6.1 Psychometric functions of the words of the PAAST in noise (Experiment 1.A)

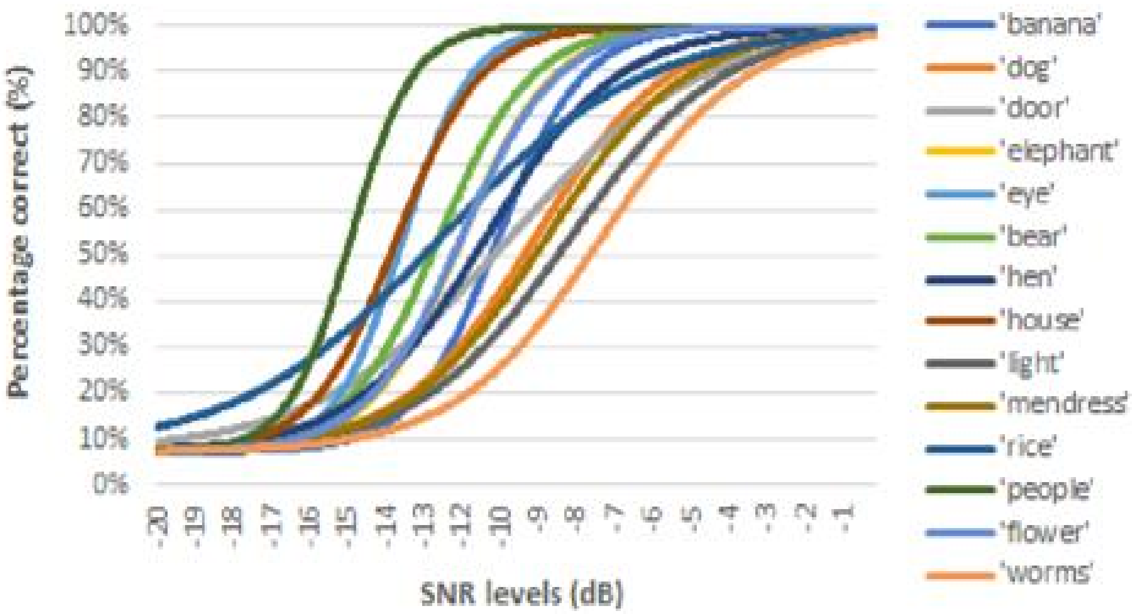
Words	Location (dB SNR)	SE	Slope	%/dB	pDev
BANANA	-12.1	0.8	0.6	16.1	1
DOG	-11.9	0.8	0.7	17.6	1
DOOR	-13.8	1.2	0.5	11.7	1
ELEPHANT	-13.5	0.8	0.7	17.8	1
EYE	-14.3	0.7	1	25.8	1
BEAR	-14.2	0.9	0.7	17.2	1
HEN	-10.9	Error	0.3	8.3	1
HOUSE	-14.1	0.8	0.8	20.6	1
LIGHT	-10.4	1.1	0.5	12	1
MENDRESS	-11.2	0.8	0.7	17.6	1

Words	Location (dB SNR)	SE	Slope	%/dB	pDev
RICE	-15.1	Error	0.2	4.6	1
PEOPLE	-15.9	0.7	1.2	29.1	0.8
FLOWER	-14.2	1	0.6	15.3	1
WORMS	-9.6	1	0.6	14.9	1
Mean	-12.9	0.9	0.7	16.3	1.0

E.6.2 Psychometric function plots of the words of the PAAST in noise (Experiment 1.A)



E.6.3 Psychometric function plots of the words of the PAAST in noise in the pre-equalisation stage in the previous study (permission to use the figure from Al-Kahtani, 2020 granted)



Appendix F Experiment 1.B: Additional data

F.1 Effect of words and repeats on the words of the PAAST SiQ

The SRTs of all the words in repeat 1 and repeat 2 for all participants were statistically analysed to study the effect of the word and repeats on the SRTs. Studentized residuals (SRE) (explained in Section 5.4.2) were calculated for each data point, revealing that all SREs were within the acceptable range (-3 to +3) concluding that there were no outliers in the data. The SRTs were considered normally distributed ($p > 0.05$) except for the words FLOWER (repeat 2) ($p = 0.027$) and PEOPLE (repeat 1) ($p = 0.04$) as assessed by Shapiro-Wilk's test of normality on the studentized residuals. it was assumed that the violation was not so severe as to warrant transformations, particularly with all other residuals being normally distributed, and considering the robustness of the ANOVA (Blanca *et al.*, 2017).

Sphericity of the data was assessed by Mauchly's test of sphericity which indicated that the assumption of sphericity was violated for the two-way interaction as well as the words' effects ($p < 0.001$). Epsilon estimate using Greenhouse-Geisser correction was used to evaluate significance of results in the test of within subject effects. The reason Greenhouse-Geisser correction was chosen was because according to Maxwell and Delaney (2004), if estimated epsilon (ε) is less than 0.75, this correction can be used. The "Repeats" within subject effect was not tested for sphericity because it consists of two categories.

Overall effect of words and repeats on SRTs

The results of the analysis showed that there was no statistically significant two-way interaction between Repeats and words, F(13, 208) = 1.02, p = 0.435 (Figure 8.1)

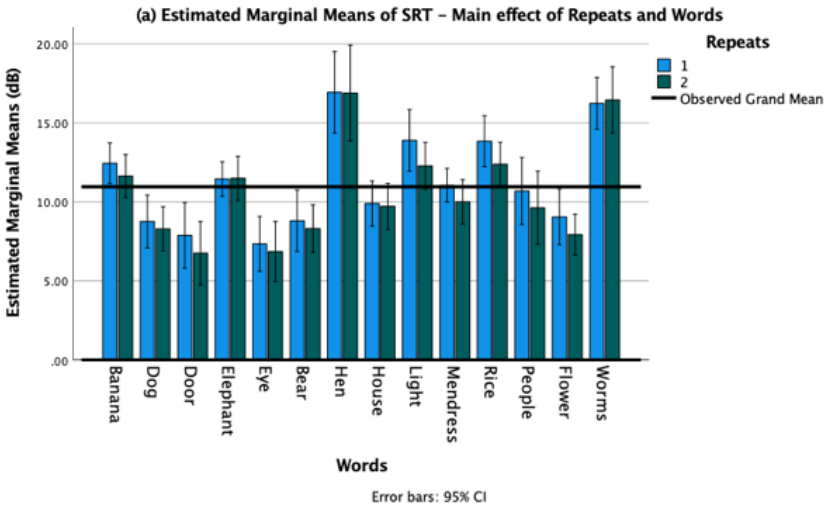


Figure 8.1 Bar chart of the estimated marginal means of the main effect of repeats and words on SRTs

Effect of words

The main effect of words showed that there was a statistically significant difference between the SRTs of the words, F(13, 208) = 24.65, p <0.001 (Figure 8.2), with a large size effect (partial-eta-squared = 0.61) (Field, 2009 (p. 57)).

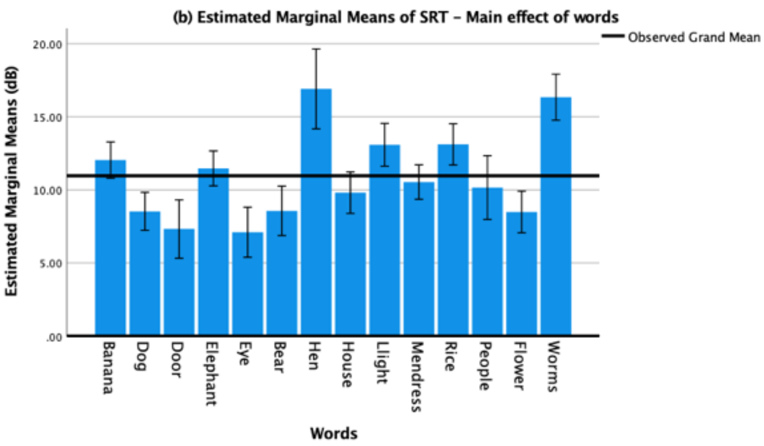


Figure 8.2 Bar chart of the estimated marginal means of the main effect of words on SRTs

The pairwise comparisons are represented in Table 8.1 indicating that the words that had significantly different SRTs than most were WORMS followed by HEN. The words with significant differences with the least number of words were PEOPLE followed by ELEPHANT.

Table 8.1 Pairwise comparison between the words of the PAAST SiQ from RM-ANOVA analysis: grey cells represent significant differences in SRTs between the compared words

	Banana	Dog	Door	Elephant	Eye	Bear	Hen	House	Light	Mendress	Rice	People	Flower	Worms
Banana														
Dog														
Door														
Elephant														
Eye														
Bear														
Hen														
House														
Light														
Mendress														
Rice														
People														
Flower														
Worms														

Effect of Repeats

The main effect of repeats showed that there was a statistically significant difference between the SRTs of the repeats, $F(1, 16) = 16.23$, $p < 0.001$, with a large size effect (partial-eta-squared = 0.5) (Feild, 2009 (p. 57)). Mean SRTs of all the words exhibited a significant decrease of 0.69 dB (SE = 0.17, 95%CI = -1.05 to -0.33) in repeat 2 (Figure 8.3).

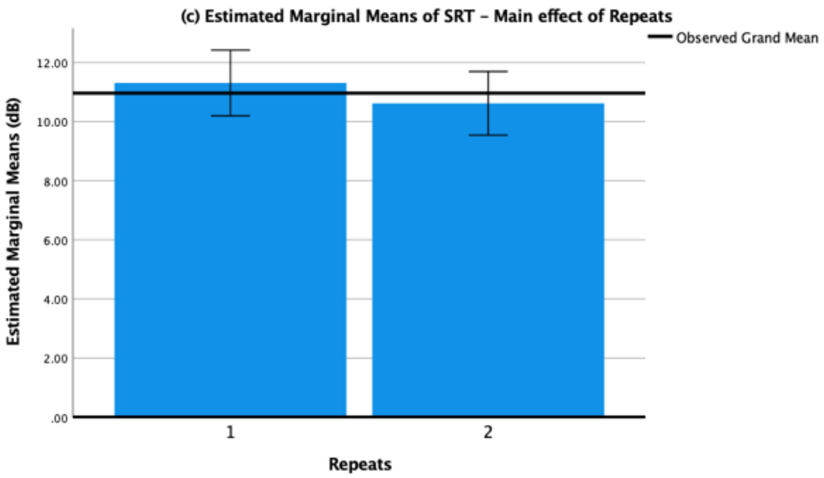


Figure 8.3 Bar chart of the estimated marginal means of the main effect of repeats on SRTs

F.2 Monte Carlo Simulations

F.2.1 Background

Assessment of speech tests should include ensuring the parameters used would lead to accurate and precise results. This is especially important when evaluating a newly developed test in a different language, such as the PAAST SiQ.

Adaptive procedure parameters that can influence the threshold include initial step size, total number of trials/ number of stimuli, number of reversals, method of calculating the threshold, and the step size rule (Zaltz *et al.*, 2019). One method of ensuring the effectiveness of parameters of any "tool" is performing computer simulations. Computer simulations use models such that for any given set of inputs or model parameters, for example, data from previous human experiment, the data is then run several hundred times, and an outcome is observed. Advantages of computer simulations include (Bonate, 2001):

1- Providing outcomes without the need to conduct an experiment
2- Shedding light on important variables
3- Providing information regarding validity of a tool unattainable from human experiments due to time and sample size constraints.
4- Answering a question of "what-if" about a tool. What if the test was performed on 200 as opposed to 400 people, or in the case of our study, what if we used a very low initial level, so close to the thresholds that in real life would be difficult to test?

A Monte Carlo simulation (MCS) is a type of simulation that relies on repeated random sampling and statistical analysis to compute the results (Raychaudhuri, 2008). They are stochastic simulations, meaning that some or all the model parameters have some degree of random variability associated with them. In order to perform MCS, the sampling distribution of the model parameters (inputs) must be defined a priori, for example the reversals and initial levels of the PAAST SiQ adaptive procedure. Monte Carlo simulation repeatedly simulates the model, each time drawing a different random set of values (inputs) from the sampling distribution of the model parameters, the result of which is a set of possible outcomes (outputs) (Bonate, 2001).

A number of studies assessing speech test parameters used computer simulations including MCS (Dingemanse and Goedegebure, 2020; Watkins et al. , 2020; Zaltz *et al.*, 2019; Tronstad, 2017).

Monte Carlo simulations of the PAAST SiQ were conducted through MATLAB using a code for interleaving adaptive track test which was considered the engine that runs the test, as well as a code file with parameters that could be changed for each condition. The reference data were a set of locations and slopes of psychometric functions obtained from participants using the MoCS from Experiment 1.A (Section 3.3).

F.2.2 Aims and Method

The main aims for conducting the simulations were:

1- To ensure that the parameters used for the human experiment are precise and accurate
2- To understand the effect of the parameters on the adaptive procedure

The parameters used for the ILAP applied in Experiment 1.B (on actual participants) are displayed in Table 8.2.

Table 8.2 Parameters used in Exp 1.B (ILAP)

initial level (IL) (dBA)	**Maximum level (dBA)**	**Reversals**	**Step size (dB)**	**Down-up rule**
30	50	1,1,6	8,4,2	1D1U

To test these parameters, simulations of different conditions were conducted to observe the resulting speech recognition threshold (SRT). These parameters included Initial Level, number of reversals, slope, and location (Table 8.3)

Table 8.3 Parameters used in the simulation

initial level (IL) (dBA)	**Maximum level (dBA)**	**Reversals (R)**	**Step size (dB)**	**Down-up rule**	**Location (L) (dBA)**	**Slope (S)**
20	50	1,1,6	8, 4, 2	1D1U	6	0.1
40		1,1,12			12	0.5
					18	0.9

The locations (L) were chosen to include the mean location (12 dB) obtained from Experiment 1 where the PAAST SiQ was applied to NH adults in MoCS, as well as the highest and lowest locations obtained. The slopes were obtained from the same experiment to include a value close to the mean slope (0.4) in addition to shallow and steep slope values. The initial level (IL) and reversals (R) values included similar value to those used in the actual experiment (20 dB and 6 reversals, respectively) as well as one more value for each. The rest of the parameters were unchanged because the main interest was on the effect of IL, reversals, location and slope on the mean SRT.

For clarification, the slope gradients 0.1, 0.5 and 0.9 are equivalent to 2.5, 12.5, and 22.5 %/dB, respectively, based on the formula mentioned in Section 3.2.3.

F.2.3 Analysis Strategy

The main observations of interest in the simulation were the accuracy (also referred to as systemic bias) and precision (Figure 8.4).

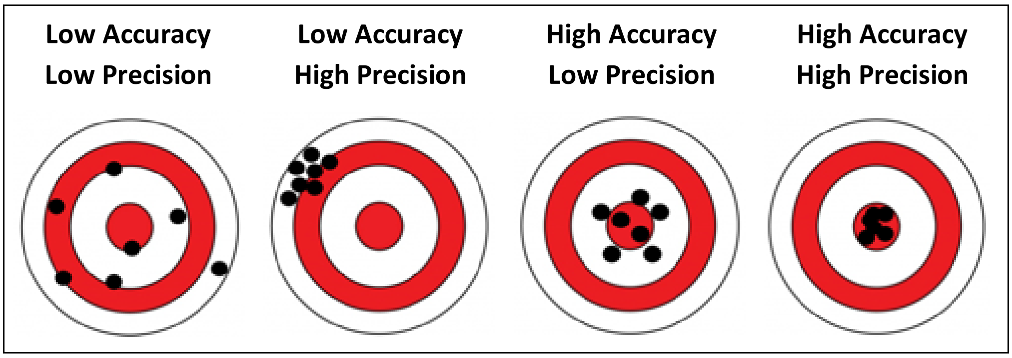


Figure 8.4 Illustration of an example of accuracy and precision (adapted from Ruotsala (2016)

Accuracy of a measurement procedure refers to how close the values of the mean SRTs are to the true expected value (Zaltz *et al.*, 2019), this could be assessed by calculating the difference between the mean SRTs and the true threshold ($V^{True} – V^{Estimate} = V^{Diff}$). The smaller the V^{Diff} the more accurate the results are. The V^{Diff} was presented as an absolute value because the interest of this study is how much the difference is between the estimated and the true thresholds. The numbers are presented without signs, and the negative numbers are presented between brackets, for clarification. Precision is how similar the mean SRT values are across all conditions, in other words, the variability of the estimate, which was assessed by the SD. The smaller the SD value, the more precise the results are.

Before discussing the results, it is important to understand the relationship between the location and the SRT because they are both points on the x-axis in the psychometric function (PF) that correspond to a certain "percentage" on the y-axis. The location is defined as the steepest point on the PF, in other words, the point of inflection (Kingdom and Prins, 2010). This point most commonly corresponds to the 50^{th} percentile (Solomon, 2011), and the accuracy of this value is determined by the function of the slope of the PF, and possibly other parameters of the test. The SRT on the other hand is the point on the PF that corresponds to a certain percentage based on the down-up rule, and since 1D1U was used, it is expected that in a closed set task, the SRT corresponds to approximately 50% (Shen, 2013). Considering these points, it is expected for the location and SRT to be approximately at the same position (50^{th} percentile). The true value at the 50^{th} percentile (True SRT) can be determined by using a code in MATLAB, this code calculates the corresponding percentile to the resulting SRT as well.

It is important to note that although 100 runs of simulations were performed, this is an in ILAP of 14 words/tracks with equal spread of location, therefore any differences between the words are eliminated rendering them equal. The mean SRTs of all runs and all tracks were calculated, yielding 1400 data points for each condition. The terms used for each condition where SRTs were obtained are listed in Table 8.4.

Table 8.4 Terms used to describe the simulated conditions and the parameters used for each condition (All R: the mean SRTs of the condition in all reversals, All IL: the mean SRTs of the condition in all Initial levels, At each slope: conditions measured at each slope, At each Location: conditions measured at each location).

Term	**Parameters tested to obtain SRTs**			
	Initial Level (IL) (dBA)	**Reversals (R)**	**Location (L) (dBA)**	**Slope (S)**
IL20R6L6	20	6	6	At each slope
IL20R6L12	20	6	12	At each slope
IL20R6L18	20	6	18	At each slope
IL20R12L6	20	12	6	At each slope
IL20R12L12	20	12	12	At each slope
IL20R12L18	20	12	18	At each slope
IL40R6L6	40	6	6	At each slope
IL40R6L12	40	6	12	At each slope
IL40R6L18	40	6	18	At each slope
IL40R12L6	40	12	6	At each slope
IL40R12L12	40	12	12	At each slope
IL40R12L18	40	12	18	At each slope
IL20L6	20	All R	6	At each slope
IL20L12	20	All R	12	At each slope
IL20L18	20	All R	18	At each slope
IL40L6	40	All R	6	At each slope
IL40L12	40	All R	12	At each slope
IL40L18	40	All R	18	At each slope
R6L6	All IL	6	6	At each slope
R6L12	All IL	6	12	At each slope
R6L18	All IL	6	18	At each slope
R12L6	All IL	12	6	At each slope

Term	Parameters tested to obtain SRTs			
	Initial Level (IL) (dBA)	**Reversals (R)**	**Location (L) (dBA)**	**Slope (S)**
R12L12	All IL	12	12	At each slope
R12L18	All IL	12	18	At each slope
I20R6S01	20	6	At each Location	0.1
IL20R6LS05	20	6	At each Location	0.5
IL20R6LS09	20	6	At each Location	0.9
IL20R12S01	20	12	At each Location	0.1
IL20R12S05	20	12	At each Location	0.5
IL20R12S09	20	12	At each Location	0.9
IL40R6S01	40	6	At each Location	0.1
IL40R6S05	40	6	At each Location	0.5
IL40R6S09	40	6	At each Location	0.9
IL40R12S01	40	12	At each Location	0.1
IL40R12S05	40	12	At each Location	0.5
IL40R12 S09	40	12	At each Location	0.9

F.2.4 Results

Accuracy

The differences between the true SRT and the estimated SRTs for each condition are displayed in Table 8.5.

Table 8.5 Mean estimated SRTs and differences (V^{Diff}) between the True SRT and the estimated SRTs in different conditions at 6 dBA Location (IL: Initial level, R: Reversals)

	Measure	Location 6			Location 12			Location 18		
		Slope 0.1	Slope 0.5	Slope 0.9	Slope 0.1	Slope 0.5	Slope 0.9	Slope 0.1	Slope 0.5	Slope 0.9
	True SRT	4.66	5.73	5.85	10.66	11.73	11.85	16.66	17.73	17.85
Total across IL and all R	**Mean SRT**	7.20	5.83	5.86	12.53	11.80	11.87	17.74	17.71	17.81
	V^{Diff}	(2.54)	(0.10)	(0.01)	(1.87)	(0.07)	(0.02)	(1.08)	0.02	0.04
IL 20	**6 Reversals**	6.41	5.84	5.82	11.17	11.84	11.94	16.08	17.63	17.82
	V^{Diff}	(1.75)	(0.11)	0.03	(0.51)	(0.11)	(0.09)	0.58	0.10	0.03
	12 Reversals	5.88	5.78	5.83	11.21	11.72	11.85	16.16	17.67	17.81
	V^{Diff}	(1.22)	(0.05)	0.02	(0.55)	0.01	0.00	0.50	0.06	0.04
IL 40	**6 Reversals**	9.07	5.93	5.95	14.41	11.88	11.88	20.04	17.80	17.78
	V^{Diff}	(4.41)	(0.20)	(0.10)	(3.75)	(0.15)	(0.03)	(3.38)	(0.07)	0.07
	12 Reversals	7.45	5.76	5.83	13.34	11.78	11.81	18.68	17.73	17.82
	V^{Diff}	(2.79)	(0.03)	0.02	(2.68)	(0.05)	0.04	(2.02)	(0.00)	0.03
IL across all R	**Avg. IL20 V^{Diff}**	(1.49)	(0.08)	0.03	(0.53)	(0.05)	(0.04)	0.54	0.08	0.04
	Avg. IL40 V^{Diff}	(3.60)	(0.11)	(0.04)	(3.21)	(0.10)	0.01	(2.70)	(0.04)	0.05
R across all IL	**Avg. R6 V^{Diff}**	(3.08)	(0.15)	(0.03)	(2.13)	(0.13)	(0.06)	(1.40)	0.02	0.05
	Avg. R12 V^{Diff}	(2.01)	(0.04)	0.02	(1.62)	(0.02)	0.02	(0.76)	0.03	0.04

Location and Slope

The mean SRTs for all locations in all conditions (regardless of the reversals and IL) where the slopes are 0.9 and 0.5 were very similar to the true value (V^{Diff} <0.2 dB) (Figure 8.5). In conditions where the slope was 0.1, the mean SRTs (averaged across reversals and ILs) differed up to (2.5) dB from the true SRT at location 6 (Table 8.5).

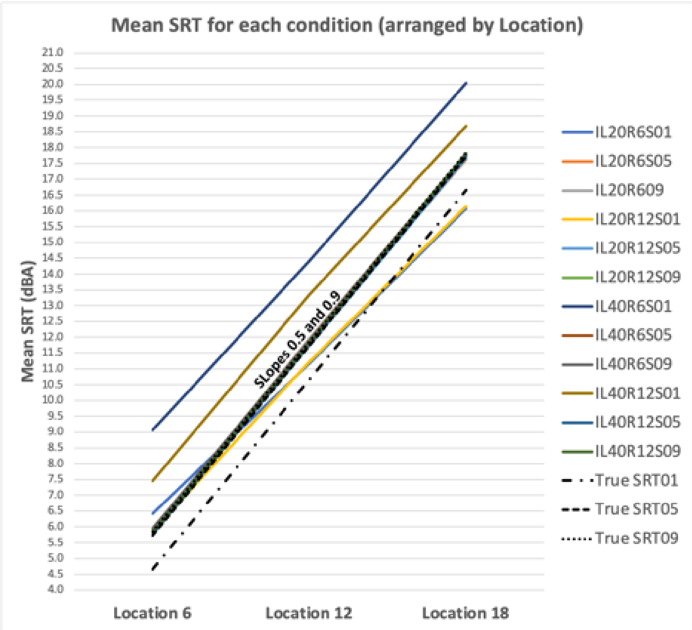


Figure 8.5 Line chart of the mean SRTs for each condition (arranged by location), where the difference of mean SRT from the true SRTs at each slope can be observed.

Although the location did not seem to influence the accuracy of the mean SRT when the slopes were 0.5 and 0.9 (V^{Diff} <0.2 dB), there was a trend towards a small increase in V^{Diff} for slope 0.5 as location increased, where the V^{Diff} in locations 6, 12, and 18 were (0.01), (0.02) and 0.04 dB, respectively. In the shallow slope (0.1) the larger the location was (across all reversals and IL) the closer the mean SRT was to the true threshold (location 18 V^{Diff} = (2.54) dB) compared to the smaller location (location 6 V^{Diff} = (1.08) dB).

Looking at the mean SRTs from the slope perspective (Figure 8.6), it could be observed that at the 0.1 slope, the locations varied up to (4.4), (3.8), and (3.4) dB from the true threshold at locations 6, 12, and 18 respectively, notably when the parameters were 6 reversals and IL was 40.

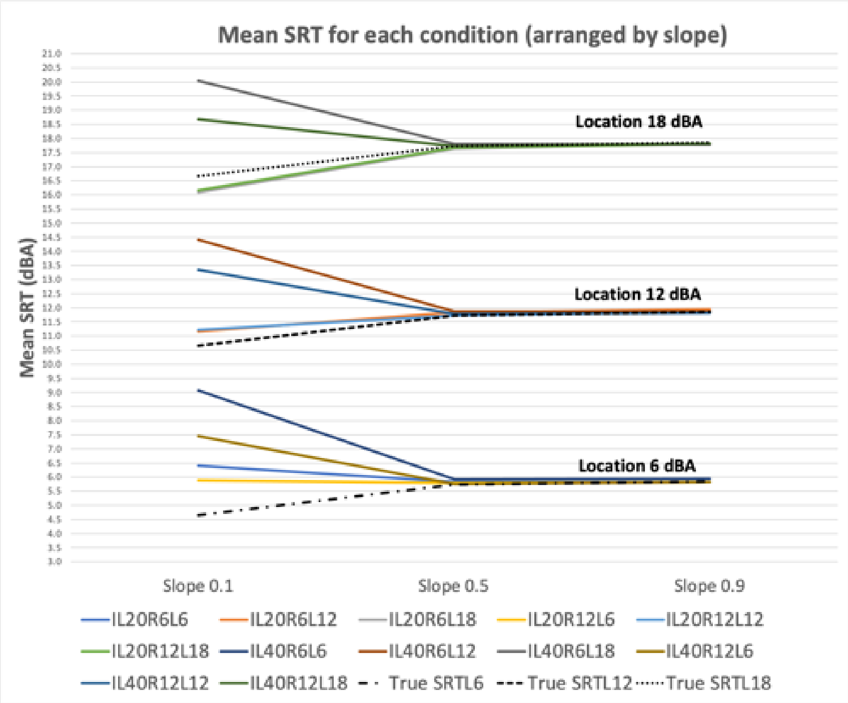


Figure 8.6 Line chart of the mean SRTs for each condition (arranged by slope), where the differences in mean SRT of each location based on slope, as well the difference of mean SRT from the true SRTs at each location can be observed

Initial Level

In the medium and steep slopes, the V^{diff} was <0.2 dB with all initial levels (in all reversals) at all locations (Figure 8.7). In the shallowest slope, the smaller the distance between Initial level and SRT (IL-SRT) the more accurate the mean SRTs were. For example, in the case of IL=20 dB and location 18 dB (IL-SRT = 2 dB), $V^{diff.}$ was 0.5 dB, on the other hand, in the case of IL = 40 and location 18 (IL-SRT = 22 dB), $V^{diff.}$ was (2.7) dB (Figure 8.7).

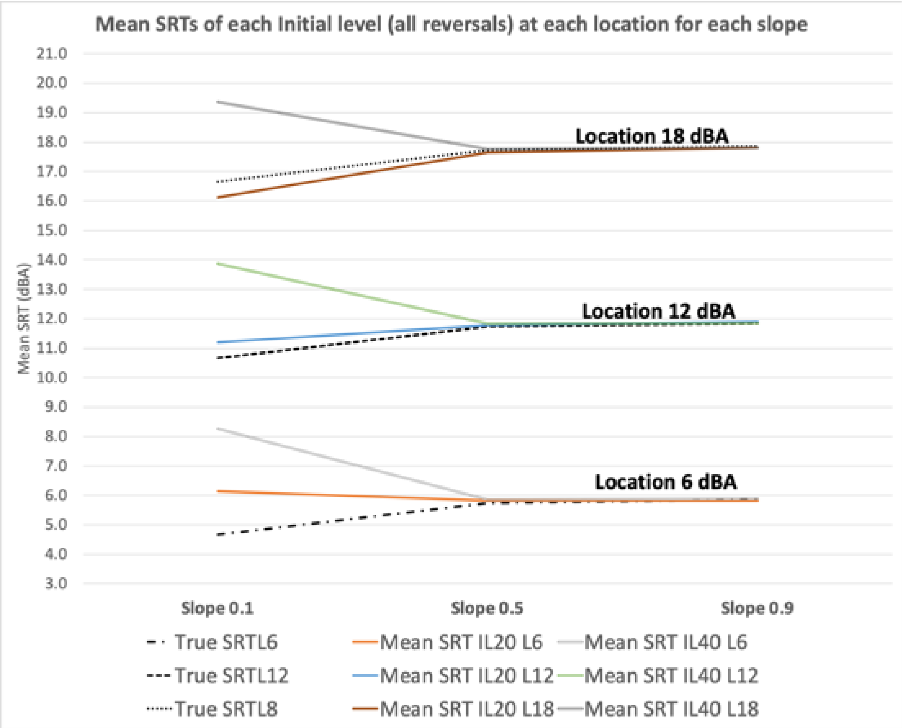


Figure 8.7 Line chart of the mean SRT for each Initial level at each location (for all reversals) arranged based on slopes, showing the differences in the mean SRT between conditions as well as the True SRT for each location.

Reversals

Similar to the initial level, V^{diff} obtained with medium and steep slopes were close to 0 with all reversals. Whereas in the shallow slopes, the V^{diff} reached up to (3.1) dB with location 6 and 6 reversals (Figure 8.8). The smallest V^{diff} in the shallow slope was seen with 12 reversals at location 18 (V^{diff} = (0.8) dB) (Table 8.5).

In the case of the shallow slope, the accuracy was decreased for all reversals and all locations (up to 2 dB difference from the true value) (Figure 8.8), but the accuracy somewhat improved by approximately 1 dB with 12 reversals compared to 6 reversals.

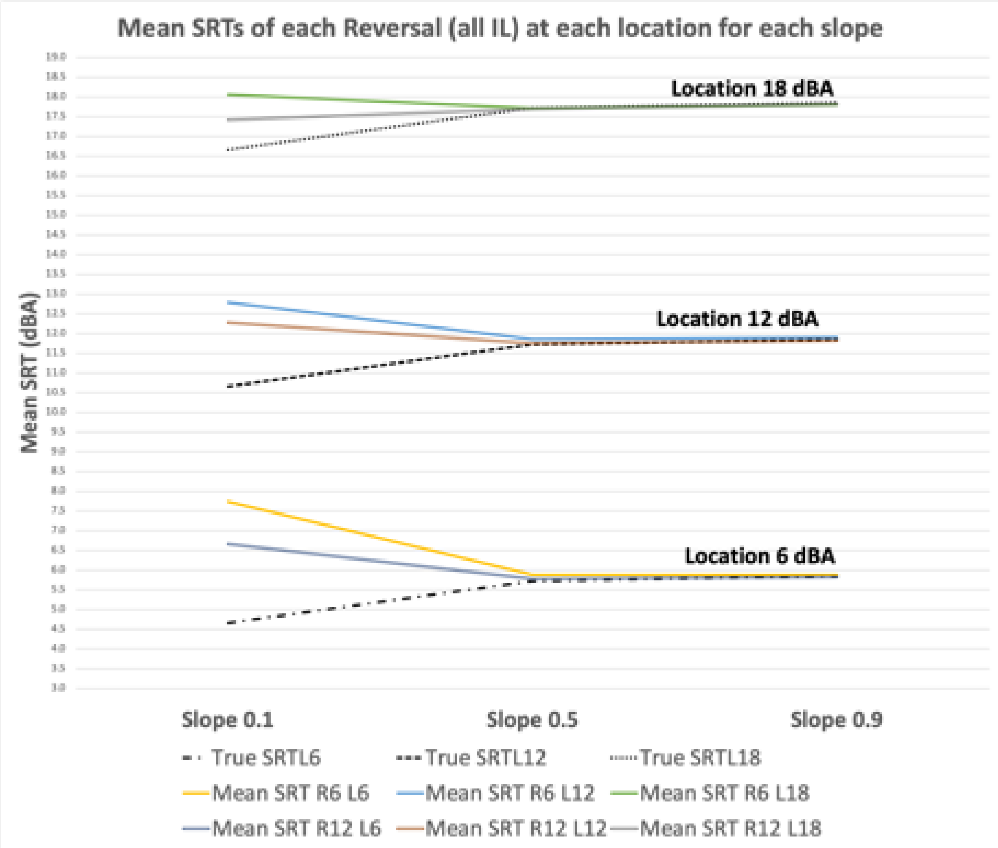


Figure 8.8 Line chart of the mean SRT for each reversal at each location (for all initial levels) arranged based on slopes, showing the differences in the mean SRT between conditions as well as the True SRT for each location.

Precision

The precision of the PAAST SIQ was assessed by the SDs in each condition. The mean estimated SRTs and SDs of each condition are displayed in Table 8.6.

Table 8.6 Mean estimated SRTs and standard deviations (SDs) for each condition at each location (IL: Initial level, R: Reversals)

	Measure (dBA)	Location 6			Location 12			Location 18		
		Slope 0.1	Slope 0.5	Slope 0.9	Slope 0.1	Slope 0.5	Slope 0.9	Slope 0.1	Slope 0.5	Slope 0.9
Total across IL and R	**Mean SRT**	7.20	5.83	5.86	12.53	11.80	11.87	17.74	17.71	17.81
	SD	4.23	1.08	0.71	4.04	1.04	0.69	3.97	1.07	0.72
Initial Level 20	**R6**	6.41	5.84	5.82	11.17	11.84	11.94	16.08	17.63	17.82
	SD	5.27	1.53	1.05	5.17	1.48	0.97	5.14	1.59	1.07
	R12	5.88	5.78	5.83	11.21	11.72	11.85	16.16	17.67	17.81
	SD	4.29	1.10	0.73	4.19	1.09	0.70	4.14	1.17	0.75
Initial Level 40	**R6**	9.07	5.93	5.95	14.41	11.88	11.88	20.04	17.80	17.78
	SD	6.56	1.62	0.99	6.00	1.54	1.04	5.90	1.49	1.06
	R12	7.45	5.76	5.83	13.34	11.78	11.81	18.68	17.73	17.82
	SD	5.04	1.14	0.75	4.85	1.09	0.75	4.67	1.10	0.72
IL across all R	**Mean SD IL20**	4.78	1.31	0.89	4.68	1.29	0.83	4.64	1.38	0.91
	Mean SD IL40	5.80	1.38	0.87	5.42	1.32	0.90	5.29	1.30	0.89
R across all IL	**Mean SD R6**	5.92	1.57	1.02	5.59	1.51	1.01	5.52	1.54	1.07
	Mean SD R12	4.66	1.12	0.74	4.52	1.09	0.73	4.41	1.14	0.73

Location, slope, and Initial Level

The SDs of the mean SRTs were small for all the conditions at slopes 0.5 and 0.9. The most precise mean SRTs (SD = 0.7 dB) were obtained when the slope was 0.9 and the reversals were 12, regardless of the initial level and location (Figure 8.9). In the shallow slope condition, the precision decreased (SD up to 6.6 in IL= 40, reversals= 6 and location= 6 (IL-SRT = 36 dB)), but improved (SD = 5.9) in location 18 (IL-SRT = 22 dB), i.e. when the IL-SRT was smaller (Table 8.6).

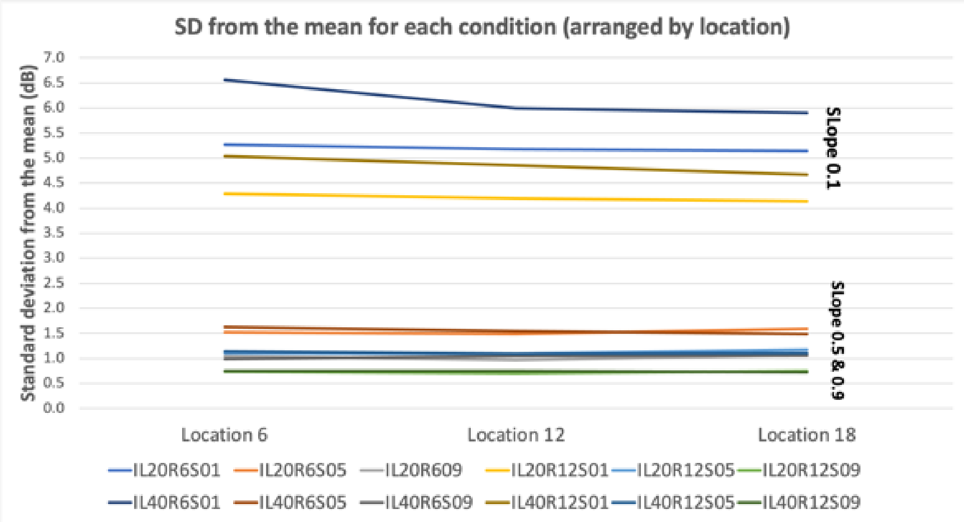


Figure 8.9 Line chart of the standard deviations from the mean SRT for each condition (arranged by location). The SDs of conditions where the slope was 0.1 were larger the conditions where the slopes were 0.5 and 0.9.

Reversals

At 12 reversals, precision improved by approximately 0.4 and 0.3 dB SD in slopes 0.5 and 0.9, respectively, regardless of the other parameters Figure 8.10). In the shallow slope the SDs vary between conditions and are generally higher (up to SD = 6.6 dB) with 6 reversals.

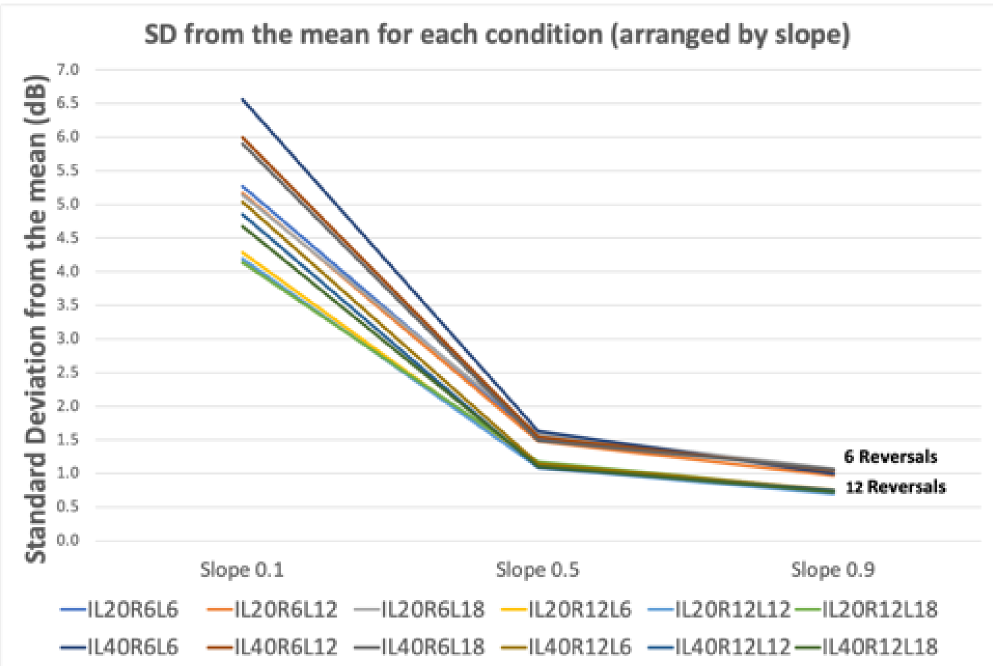


Figure 8.10 Line chart of the SDs from the mean SRT for each condition (arranged by slope). The SDs were smallest at conditions where the slope was 0.9, especially at reversal = 12.

Appendix G OM-6 Translation

G.1 Approval of the author of the OM-6 to translate the OM-6

On Tue, Dec 25, 2018 at 1:37 PM +0300, "Richard Rosenfeld" <rich@richrosenfeld.com> wrote:

> Hi Sarah:
> You are welcome the use the OM6 as you please. The only requirement is to cite the supporting research articles in any publication(s) that results from your research.
> Best regards,
> Rich Rosenfeld
>
> Richard M. Rosenfeld, MD, MPH, MBA
> Distinguished Professor and Chairman,
> Department of Otolaryngology
> President of the Medical and Dental Staff
> SUNY Downstate Medical Center
> Brooklyn, NY, USA
> 718-270-1638 office
> 718-637-1537 cell

G.2 OM-6 English

The English (original) OM-6

Quality-of-life Questionnaire (Arabic OM-6) for children with Otitis Media with Effusion (OME)

Instructions: Please help us understand the impact of ear infections or fluids on your child's quality of life by checking one box [X] for each question below. Thank you

Physical suffering: Ear pain, Ear discomfort, Ear discharge, ruptured eardrum, high fever, or poor balance. How much of a problem for your child during the past four weeks?

[ ] Not present/ no problem	[ ] Hardly a problem at all	[ ] Quite a bit of a problem
	[ ] Somewhat of a problem	[ ] Very much a problem
	[ ] Moderate problem	[ ] Extreme problem

Hearing loss: Difficulty hearing, questions must be repeated, frequently says "what?", television is excessively loud. How much of a problem is for your child during the past four weeks?

[ ] Not present/ no problem	[ ] Hardly a problem at all	[ ] Quite a bit of a problem
	[ ] Somewhat of a problem	[ ] Very much a problem
	[ ] Moderate problem	[ ] Extreme problem

Speech impairment: Delayed speech, poor pronunciation**,** difficult to understand, or are unable to repeat words clearly. How much of a problem for your child during the past four weeks?

[ ] Not present/no problem	[ ] Hardly a problem at all	[ ] Quite a bit of a problem
(or not applicable)	[ ] Somewhat of a problem	[ ] Very much a problem
	[ ] Moderate problem	[ ] Extreme problem

Emotional distress: irritable, frustrated, sad, restless, or poor appetite. How much of a problem for your child during the past four weeks as a result of ear infections or fluid?

[ ] Not present/ no problem	[ ] Hardly a problem at all	[ ] Quite a bit of a problem
	[ ] Somewhat of a problem	[ ] Very much a problem
	[ ] Moderate problem	[ ] Extreme problem

Activity Limitations: playing, sleeping, doing things with friends/family, attending school or daycare. How limited have your child's activities been doing the past four weeks because of ear infections or fluids?

[ ] Not limited at all	[ ] Hardly limited at all	[ ] Moderately limited
	[ ] They're slightly limited	[ ] Very limited
	[ ] Slightly limited	[ ] Severely limited

Caregiver concerns: how often have you, as a caregiver, been worried, concerned, or inconvenienced because of your child's ear infections or fluid over the past four weeks?

[ ] None of the time	[ ] Hardly any time	[ ] A good part of the time
	[ ] A small part of the time	[ ] Most of the time
	[ ] Some of the time	[ ] All of the time

Overall, how would you rate your child's quality-of-life as a result of ear infections or fluid?
(circle one number)

0 1 2 3 4 5 6 7 8 9 10

Worse Possible Quality-of-Life — Half-way between Worst and Best — Best possible Quality-of-Life

G.3 Arabic OM-6

استبيان جودة الحياة لدى الأطفال المصابين بالتهابات الأذن النزمنة و المتكررة (OM-6 العربي) - V2

توجيهات: نرجو منكم مساعدتنا على فهم تأثير التهابات أو سوائل الأذن على جودة حياة أبنكم/ابنتكم وذلك بوضع علامة [X] على أحد خيارات الأسئلة أدناه. شكرا لكم.

المعاناة الجسدية: الام في الأذن أو عدم ارتياح في منطقة الأذن أو إفرازات أذنية أو حمى مرتفعة أو فقدان التوازن. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

فقدان السمع: صعوبة في السمع، الاضطرار لتكرار الأسئلة، قول كلمة "ماذا؟" مرات متعددة، رفع صوت التلفاز. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

اضطراب في الكلام: تأخر في الكلام، نطق ضعيف، صعوبة في الفهم، عدم القدرة على تكرار الكلمات بوضوح. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

الاضطراب العاطفي: سرعة الغضب، العصبية، الحزن، عدم الارتياح، فقدان الشهية. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

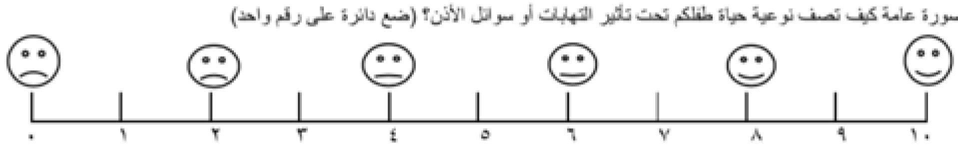
الحد من الأنشطة: اللعب، النوم، التفاعل مع الأسرة/الأصدقاء، حضور المدرسة أو مركز العناية النهارية (الحضانة). ما درجة تأثير التهاب الأذن من ناحية الحد من أنشطة طفلكم خلال الأسابيع الأربعة الماضية؟

[ ] لم تكن محدودة أبدا
[ ] تكاد ان تكون غير محدودة
[ ] كانت محدودة لحد ما
[ ] كانت محدودة قليلا
[ ] كانت محدودة بقدر متوسط
[ ] كانت محدودة كثيرا
[ ] كانت محدودة بالكامل

قلق مقدم الرعاية: بصفتكم مقدمي الرعاية، ما مدى قلقكم / تضايقكم على طفلكم بسبب التهاب الأذن خلال الأسابيع الأربعة الماضية؟

[ ] لم أقلق أبدا
[ ] لم أقلق إلا نادرا
[ ] قلقت أحياناً قليلة
[ ] قلقت أحيانا
[ ] قلقت أحيانا عديدة
[ ] قلقت في غالب الأحيان
[ ] قلقت باستمرار

بصورة عامة كيف تصف نوعية حياة طفلكم تحت تأثير التهابات أو سوائل الأذن؟ (ضع دائرة على رقم واحد)



أفضل جودة حياة ممكنة
جودة حياة متوسطة
أسوأ جودة حياة ممكنة

الدراسة الاستقصائية الصحية المتعلقة بنوعية الحياة (OM-6) لالتهاب الأذن المزمن والمتكرر.

G.4 AOM6 form (Experiment 3)

استبيان جودة الحياة لدى الأطفال المصابين بالتهابات الأذن النزمنة و المتكررة (OM-6 العربي)

تجرى هذه الدراسة ضمن متطلبات بحث رسالة طالبة الدكتوراه د. سارة زهير السباعي من جامعة ساوثهامبتون في بريطانيا بالتعاون مع جامعة الملك عبد العزيز في السعودية و التي تتضمن دراسة جودة الحياة لدى الأطفال من عمر ٦ أشهر إلى ١٦ سنة المصابين بالتهابات الأذن الوسطى المزمنة أو المتكررة.

تعتبر دراسة جودة الحياة لدى الأطفال المصابين بالتهابات الأذن الوسطى مهمة لمعرفة مدى تأثير هذه الالتهابات على حياة الطفل من الناحية الصحية، السمعية، اللفظية، و كذلك السلوكية، و ذلك لمعرفة مدى تحسن الطفل مع العلاج .

نرجو منكم المساهمة بتطوير استبيان OM-6 العربي لقياس جودة الحياة لدى الاطفال المصابين بالتهابات الأذن الوسطى عن طريق الاجابة على الاستبيان من قبل أحد مقدمي الرعاية للطفل المصاب (الأم، الأب، الجدة أو غيره)

سيتم التعامل مع المعلومات بسرية تامة حيث لن يطلع عليها سوى الباحثة الرئيسية. شاكرين لكم حسن تعاونكم

<u>الإقرار و الموافقة على المشاركة</u>

☐ أقر بالموافقة على المشاركة في البحث و على استخدام البيانات الشخصية لغرض البحث فقط

<u>المعلومات الشخصية:</u>

تاريخ ولادة الطفل (ميلادي أو هجري) (اليوم، الشهر، السنة): / /

الجنس: [ ] ذكر

[ ] أنثى

<u>حالة الأذن</u>

الأسئلة في هذا القسم تعتمد على ما تم ابلاغكم به من قبل الطبيب، في حالة عدم المعرفة بذلك ارجو اختيار (غير متأكد)

ما نوع التهاب الأذن لدى طفلك؟

[ ] التهاب أذن مزمن مصاحب له سوائل

[ ] التهاب أذن حاد متكرر (ألتهاب حاد في الأذن أكثر من ٣ مرات خلال السنة)

[ ] غير متأكد

أي الأذنين مصابة بالالتهاب؟

[ ] الأذن اليمنى

[ ] الأذن اليسرى

[ ] كلتا الأذنين اليمنى و اليسرى

[ ] غير متأكد

هل يشكو طفلكم من أي مشاكل صحية مزمنة أخرى؟

[ ] نعم

[ ] لا

الدراسة الاستقصائية الصحية المتعلقة بنوعية الحياة (OM-6) لدى الأطفال المصابين بالتهاب الأذن الوسطى المزمن والمتكرر.

استبيان OM-6 العربي

توجيهات: نرجو منكم مساعدتنا على فهم تأثير التهابات أو سوائل الأذن على جودة حياة أبنكم/ابنتكم وذلك بوضع علامة [X] على أحد خيارات الأسئلة أدناه. شكرا لكم.

المعاناة الجسدية: آلام في الأذن أو عدم ارتياح في منطقة الأذن أو إفرازات أذنية أو حمى مرتفعة أو فقدان التوازن. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

فقدان السمع: صعوبة في السمع، الاضطرار لتكرار الأسئلة، قول كلمة "ماذا؟" مرات متعددة، رفع صوت التلفاز. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

اضطراب في الكلام: تأخر في الكلام، نطق ضعيف، صعوبة في الفهم، عدم القدرة على تكرار الكلمات بوضوح. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

الاضطراب العاطفي: سرعة الغضب، العصبية، الحزن، عدم الارتياح، فقدان الشهية. كيف أثرت هذه الأعراض على طفلكم في الأسابيع الأربعة الماضية؟

[ ] غير موجودة / لم تشكل مشكلة
[ ] تكاد ان تكون غير موجودة
[ ] مشكلة بسيطة
[ ] مشكلة متوسطة
[ ] مشكلة مقلقة لحد ما
[ ] مشكلة كبيرة
[ ] مشكلة كبيرة جدا

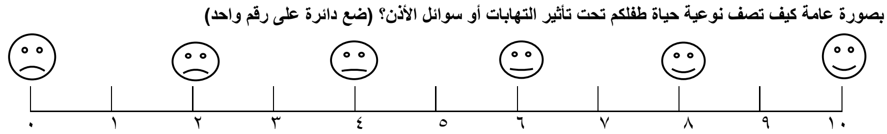
الحد من الأنشطة: اللعب، النوم، التفاعل مع الأسرة/الأصدقاء، حضور المدرسة أو مركز العناية النهارية (الحضانة). ما درجة تأثير التهاب الأذن من ناحية الحد من أنشطة طفلكم خلال الأسابيع الأربعة الماضية؟

[ ] لم تكن محدودة أبدا
[ ] تكاد ان تكون غير محدودة
[ ] كانت محدودة قليلا
[ ] كانت محدودة لحد ما
[ ] كانت محدودة بقدر متوسط
[ ] كانت محدودة كثيرا
[ ] كانت محدودة بالكامل

قلق مقدم الرعاية: بصفتكم مقدمي الرعاية، ما مدى قلقكم / تضايقكم على طفلكم بسبب التهاب الأذن خلال الأسابيع الأربعة الماضية؟

[ ] لم أقلق أبدا
[ ] لم أقلق إلا نادرا
[ ] قلقت أحياناً قليلة
[ ] قلقت أحيانا
[ ] قلقت أحيانا عديدة
[ ] قلقت في غالب الأحيان
[ ] قلقت باستمرار

بصورة عامة كيف تصف نوعية حياة طفلكم تحت تأثير التهابات أو سوائل الأذن؟ (ضع دائرة على رقم واحد)



أسوأ جودة حياة ممكنة — جودة حياة متوسطة — أفضل جودة حياة ممكنة

الدراسة الاستقصائية الصحية المتعلقة بنوعية الحياة (OM-6) لدى الأطفال المصابين بالتهاب الأذن الوسطى المزمن والمتكرر.

<u>الفحص الميداني لاستبيان OM-6 العربي</u>

بناء على استبيان جودة الحياة OM-6 باللغة العربية، الرجاء الاجابة على الاسئلة التالية المتعلقة بمدى وضوح التوجيهات، الأسئلة والأجوبة وملائمتها للمجتمع

هل توافق على أن التوجيهات، الأسئلة والأجوبة واضحة؟

[ ] أوافق بشدة [ ] أوافق [ ] محايد
[ ] لا أوافق [ ] لا أوافق بشدة

هل توافق على أن الكلمات المستخدمة في التوجيهات، الأسئلة والأجوبة سهلة ومفهومة؟

[ ] أوافق بشدة [ ] أوافق [ ] محايد
[ ] لا أوافق [ ] لا أوافق بشدة

هل توافق على أن الكلمات المستخدمة في التوجيهات، الأسئلة والأجوبة ملائمة لمجتمعنا وغير مزعجة؟

[ ] أوافق بشدة [ ] أوافق [ ] محايد
[ ] لا أوافق [ ] لا أوافق بشدة

إذا كان لديك أي تعليق على استبيان OM-6 العربي الرجاء كتابته هنا، في حال عدم وجود أي تعليق من الممكن ترك هذا الجزء خالياً

انتهت الأسئلة ... شكراً لكم على حسن تعاونكم

الدراسة الاستقصائية الصحية المتعلقة بنوعية الحياة (OM-6) لدى الأطفال المصابين بالتهاب الأذن الوسطى المزمن والمتكرر.

Appendix H Additional Analysis

H.1 The statistical significance of adding an interaction term (age)

The table below (Table F.1) shows the statistical significance of adding the interaction term β_3 (age) to the model. This was assessed using the change in "-2log(likelihood)" for the model with and without β_3, which can be tested using a chi-square distribution. The differences between the two models in "- 2log likelihood" following the chi-square distribution with one degree of freedom is 3.84. There was no significant difference between the models with or without interaction of β_3.

Table F.1 The results of -2log likelihood difference and its significance when compared to one degree of freedom

Model	**With interaction model**	**Without interaction model**	**Difference between the two models**
Change in PTA-HL (dBHL) with SRT (dB A) (β_4)	-569.814	-572.576	2.76

H.2 The difference between two correlations

An Analysis was conducted to examine the difference between the correlation coefficients based on the method by Meng, Rosenthal and Rubin (1992). The correlation coefficients examined were of the following relationships:

(1) AOM6 (Total) and SRT (dBA): (r = 0.43)
(2) AOM6 (Total) and PTA-HL (dBHL): (r = 0.33)

The analysis was conducted based on the relationship between SRT and PTA (r = 0.74). The results of the analysis showed no significant difference between the two correlations (Table F.2).

Table F.2 The difference between the correlation coefficient for AOM6 (Total) and SRT and AOM6 (Total) and PTA-HL

Difference between correlations	**95% CI of the difference**		**One-tailed p-value**	**Two-tailed p-value**
	Lower	**upper**		
0.10	-0.10	0.33	0.15	0.30

Appendix I Ethical approvals

I.1 Ethical approval (ERGO II) for Pilot Experiment

UNIVERSITY OF Southampton

ERGO II – Ethics and Research Governance Online https://www.ergo2.soton.ac.uk

Submission ID: 46958
Submission Title: Testing the Sensitivity of the Arabic to Test in Quiet to Simulated Conductive Hearing Loss Caused by Otitis Media with Effusion
Submitter Name: Sarah Alsebai

Your submission has now been approved by the Faculty Ethics Committee. You can begin your research unless you are still awaiting any other reviews or conditions of your approval.

Comments:

- Thank you for the updates. There is no little ethical risk, so I am happy to approve the application.

Click here to view the submission

I.2 Ethical Approval (KAAU) for Pilot, Experiment 1.A and 1.B

KINGDOM OF SAUDI ARABIA
Ministry 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 Committee

Initial Approval

From: Professor. Hasan Alzahrani

TO: Principal Investigator and Local Supervisor: Dr.Afaf Bamani (Otolaryngology-Audiology Unit)

First-Investigator:Dr.Sarah Alsebai (Joint study programme student with University of Southampton)

External-Supervisor:Dr.Daniel Rowan

Date: Tuesday, November 13, 2018

CC:Vice-Dean, University /Hospital Director & Academic Affairs & File & Monitoring Committee

RE:" Validation of The Arabic Toy Test in Quiet" (Reference No 569-18) Non-Intervention Cross Sectional.

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

- Application for Research Unit of Biomedical Ethics. KAU FoM Form.
- Informed consent V Arabic ,English
- Study Protocol .

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

1. The PI and Supervisors are responsible to get Academic Affairs, hospital and departmental approval. , according to bylaws they must get the administrative approval from organization collaborators outside KAUH.
2. The approval of conduct of this study will be automatically suspended after 03 months in case if no submission of " Continuing Review Progress Report KAU FoM Forms " to review by REC, Monitoring Committee.
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.

Mohammed S. aLsearee (Reference No 569-18)

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

فاكس ٦٤٠٠٥٩٢ / ٦٤٠٨٤٥١
Fax : 6400592/ 6408451

٦٩٥٢٠٦٣ / ٦٩٥٢٤٤١
: 6952063 / 6952063

I.3 Ethics Approval (ERGO II) for Experiment 1.A and 1.B

UNIVERSITY OF Southampton

ERGO II – Ethics and Research Governance Online https://www.ergo2.soton.ac.uk

Submission ID: 46958.A1
Submission Title: Testing the Sensitivity of the Arabic Toy Test in Quiet to Simulated Conductive Hearing Loss Caused by Otitis Media with Effusion (Amendment 1)
Submitter Name: Sarah Alsebai

Your submission has now been approved by the Faculty Ethics Committee. You can begin your research unless you are still awaiting any other reviews or conditions of your approval.

Comments:

- Thank you for making the changes in this amendment so clear.

  Apologies for the slight delay on my side. I am happy there is little ethical risk and so will approve.

Click here to view the submission

I.4 Ethical Approval (ERGO II) for Experiment 2

UNIVERSITY OF Southampton

ERGO II – Ethics and Research Governance Online https://www.ergo2.soton.ac.uk

Submission ID: 46958.A2
Submission Title: Testing the Sensitivity of the Arabic Toy Test in Quiet to Simulated Conductive Hearing Loss Caused by Otitis Media with Effusion (Amendment 2)
Submitter Name: Sarah Alsebai

Your submission has now been approved by the Faculty Ethics Committee. You can begin your research unless you are still awaiting any other reviews or conditions of your approval.

Comments:

- Thank you for the updates; and especially highlighted the changes you have made - it makes it so much easier.

  I will approve on the basis:

  - that you follow all UK Government guidelines on activities during COVID
  - that you follow all local guidelines in the country where the research is being carried out on activities during COVID

  Good luck! And stay safe

I.5 Ethical approval (ERGO II) for Experiment 3

UNIVERSITY OF Southampton

ERGO II – Ethics and Research Governance Online https://www.ergo2.soton.ac.uk

Submission ID: 52801.A1
Submission Title: Developing an Arabic Quality-of-life Questionnaire (Arabic OM-6) for children with Chronic and Recurrent Otitis Media (Amendment 1)
Submitter Name: Sarah Alsebai

Your submission has now been approved by the Faculty Ethics Committee. You can begin your research unless you are still awaiting any other reviews or conditions of your approval.

Comments:

-

Click here to view the submission

TId: 23011_Email_to_submitter___Approval_from_Faculty_Ethics_committee__cat_B___C_ Id: 335297
S.Z.Alsebai@soton.ac.uk coordinator

I.6 Ethical approval (KAAU) for Experiment 3

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)
From: Professor. Hasan Alzahrani
First-investigator: Sarah Alsebai
External-Supervisor: Daniel Rowan
Date: Tuesday, December 31, 2019
CC:Vice-Dean, University /Hospital Director & Academic Affairs & File & Mentoring Committee

RE: Developing an Arabic Quality-of-life Questionnaire (Arabic OM-6) For children with chronic and Recurrent Otitis media. (Reference No 755-19) Qualitative 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 755-19)

ص.ب: ٨٠٢٠٥ جـدة: ٢١٥٨٩
P.O.Box: 80205 Jeddah: 21589
فاكس: ٦٤٠٠٥٩٢ / ٦٤٠٨٤٥١
Fax.: 6408451 /6400592
☎: ٦٩٥٢٠٦٣ / ٦٩٥٢٤٤٦
☎: 6952446 / 6952063

I.7 Ethical Approval (ERGO II) for Experiment 4

Southampton UNIVERSITY OF

ERGO II – Ethics and Research Governance Online https://www.ergo2.soton.ac.uk

Submission ID: 68934
Submission Title: Speech recognition and quality of life in Arabic children with OME and the correlation between them
Submitter Name: Sarah Alsebai

Your submission has now been approved by the Faculty Ethics Committee. You can begin your research unless you are still awaiting any other reviews or conditions of your approval.

Comments:

-
-
- Thank you for taking the time to address the points raised. Wishing you best of luck with your investigation.

Click here to view the submission

TId: 23011_Email_to_submitter___Approval_from_Faculty_Ethics_committee__cat_B___C_ Id: 441141

S.Z.Alsebai@soton.ac.uk coordinator

I.8 Ethical approval (KAAU) for Experiment 4

KINGDOM OF SAUDI ARABIA
Ministry 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 Committee

TO: Principal Investigator & Local Supervisor: Prof. Afaf Bamanie **from: Professor. Hasan Alzahrani**
(Otolaryngology and Head and Neck surgery- Audiology Unit, KAUH)

External &Main- Supervisor: Dr. Daniel Rowan

First-Investigator: Dr. Sarah Al-Sebai (Audiovestibular Medicine demonstrator/ Joint study programme student with the University of Southampton – Audiology)

Date: Tuesday, December 14, 2021 CC: Vice-Dean, University /Hospital Director & Academic Affairs & File & Monitoring Committee

RE: " Measuring Speech Recognition and Quality of life in Arabic children with OME using PAAST SiQ and Arabic OM-6 Respectively"

(Reference No 578-21) Non- Intervention (Cross sectional)

Application for Research Unit of Biomedical Ethics. KAU FoM, KAUH Form.
Data Collection Sheet

1. The PI and Supervisors are responsible to get Academic Affairs, hospital and departmental approvals, according to bylaws they must get the administrative approval from organization collaborators outside KAUH.
2. The approval of conduct of this study will be automatically suspended after 06 months in case if no submission of " Continuing Review Progress Report KAU FoM Forms " to review by REC, Monitoring Committee.
3. The Investigators will conduct the study under the direct supervision of **Prof. Afaf Bamanie**
4. 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.
5. The PI is expected to submit a final report at the end of the study.
6. The PI must provide to REC a conclusion abstract and the manuscript before publication.
7. To follow all regulations issued by the National Committee of Bio & Med ethics - 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 and we confirm that the PI is not part of the ethics committee.
The committee is fully compliant with the regulations as they relate to Ethics Committees and the conditions and principles of good clinical practice.
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

c/o

(HA-02-J-008) No of Registration At National Committee of Bio.& Med. Ethics.
Eman A.jehini (Reference No 578-21)

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

List of References

Abdulhaq, N. M. A. (2006) *Speech Perception test for Jordanian Arabic Speaking Children*, *A Dissertation part of Doctor of Phyilosophy*. University of Florida.

Adelman, C. *et al.* (2015) 'Air Conduction, Bone Conduction, and Soft Tissue Conduction Audiograms in Normal Hearing and Simulated Hearing Losses', *Journal of the American Academy of Audiology*, 26(1), pp. 101–108. doi: 10.3766/jaaa.26.1.11.

Al-Abduljawad, K. A. and Zakzouk, S. M. (2003) 'The prevalence of sensorineural hearing loss among Saudi children', *International Congress Series*, 1240(C), pp. 199–204. doi: 10.1016/S0531-5131(03)00913-0.

Al-arfaj, H. (2018) *Test-Retest Reliability of the Arabic Version of the McCormick Toy Test in a Quiet Setting*. University of Southampton.

Al-humaid, H. *et al.* (2014) 'Prevalence and risk factors of otitis media with effusion among school children in Saudi Arabia', *Otolaryngology - Head and Neck Surgery (United States)*, 151(1), p. P107. doi: 10.1177/0194599814541627a246.

Al-Kahtani, R. (2020) *Assessment of Sensorineural Hearing Loss in Children in the Kingdom of Saudi Arabia*. University of Southampton.

Al-Muhaimeed, H. S. (1996) 'Hearing impairment among "at risk" children.', *International journal of pediatric otorhinolaryngology*, 34(1–2), pp. 75–85.

Al-Quaiz, A.-J. M. (2001) 'Correlates of various presentation modes of acute otitis media in Saudi children', *Journal of Family Community Med*, 8(2), pp. 17–24.

Al-Rowaily, M. A. *et al.* (2012) 'Hearing impairments among Saudi preschool children', *International Journal of Pediatric Otorhinolaryngology*. Elsevier Ireland Ltd, 76(11), pp. 1674–1677. doi: 10.1016/j.ijporl.2012.08.004.

Alakeely, M. H. *et al.* (2022) 'The Ability of Saudi Parents' To Detect Early Language Delay in Their Children: A Study in Primary Health Care Centers, King Abdulaziz Medical City, Riyadh, Saudi Arabia', *Cureus*, 14(1), pp. 1–10. doi: 10.7759/cureus.21448.

Aldè, M. *et al.* (2021) 'Effects of COVID-19 Lockdown on Otitis Media With Effusion in Children: Future Therapeutic Implications', *Otolaryngology - Head and Neck Surgery (United States)*, 165(5), pp. 710–715. doi: 10.1177/0194599820987458.

Alkahtani, R. *et al.* (2019) 'Age of identification of sensorineural hearing loss and Characteristics of affected children: Findings from two cross-sectional studies in Saudi Arabia', *International Journal of Pediatric Otorhinolaryngology*, 122(November 2018), pp. 27–34. doi: 10.1016/j.ijporl.2019.03.019.

Allen, D. Z. *et al.* (2022) 'Impact of COVID-19 on nationwide pediatric otolaryngology: Otitis media and myringotomy tube trends', *American Journal of Otolaryngology - Head and Neck Medicine and Surgery*. Elsevier Inc., 43(2), p. 103369. doi: 10.1016/j.amjoto.2021.103369.

Alqahtani, Z. *et al.* (2017) 'Otitis Media in Children at Riyadh Capital City of KSA', *EC Microbiology*, 12(5), pp. 224–231.

Alsuhaibani, M. *et al.* (2020) 'Awareness and attitudes of Saudi parents toward otitis media in children', *Journal of Family Medicine and Primary Care*, 9(12), pp. 6177–6182. doi: 10.4103/jfmpc.jfmpc.

Ambrosio, A. and Brigger, M. T. (2014) 'Surgery for otitis media in a universal health care model: Socioeconomic status and race/ethnicity effects', *Otolaryngology - Head and Neck Surgery (United States)*. SAGE Publications Inc., 151(1), pp. 137–141. doi: 10.1177/0194599814525570.

de Andrade, K. C. L. *et al.* (2013) 'Non-flat audiograms in sensorineural hearing loss and speech perception', *Clinics*, 68(6), pp. 815–819. doi: 10.6061/clinics/2013(06)15.

Ashoor, A. A. and Fuer, F. (2013) 'Management of otitis media with effusion in children', *Bahrain Medical Bulletin*, 35(3), pp. 123–125. Available at: http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed7&NEWS=N&AN=16363264.

Ashoor, A. A. and Prochazka, T. (1982) 'Saudi Arabic Speech Audiometry', *Audiology*, 21(6), pp. 493–508. doi: 10.3109/00206098209072761.

Atkinson, H., Wallis, S. and Coatesworth, A. P. (2015) 'Otitis media with effusion', *Postgraduate Medicine*, 127(4), pp. 381–385. doi: 10.1080/00325481.2015.1028317.

BAA (2013) 'Recommended Procedure. Tympanometry', *British Society of Audiology*, (August), pp. 1–20. Available at: http://www.thebsa.org.uk/wp-content/uploads/2014/04/BSA_RP_Tymp_Final_21Aug13_Final.pdf.

Beaton, D. E. *et al.* (1999) 'Guidelines for the Process of Cross-Cultural Adaptation of Self-Report Measures', *Acta Odontologica Scandinavica*, 57(4), pp. 225–230. doi: 10.1080/000163599428823.

Bell, S. (2001) 'Good Practice Guide No. 11 - Introductory Guide to Uncertainty of Measurement',

Measurement Good Practice Guide, (2), p. 41.

Benner, A. D. and Mistry, R. S. (2020) 'Child Development During the COVID-19 Pandemic Through a Life Course Theory Lens', *Child Development Perspectives*, 14(4), pp. 236–243. doi: 10.1111/cdep.12387.

Bhat, A. (2021) 'Analysis of the SPARK study COVID-19 parent survey: Early impact of the pandemic on access to services, child/parent mental health, and benefits of online services', *Autism Research*, 14(11), pp. 2454–2470. doi: 10.1002/aur.2618.

Blanca, M. J. *et al.* (2017) 'Non-normal data: Is ANOVA still a valid option?', *Psicothema*, 29(4), pp. 552–557. doi: 10.7334/psicothema2016.383.

Bland, J. M. and Altman, D. G. (1996) 'Measurement Error', *BMJ*, 312, pp. 275–286. doi: 10.1002/9781118763025.ch27.

Blank, S. J. *et al.* (2014) 'Caregiver quality of life is related to severity of otitis media in children', *Otolaryngology - Head and Neck Surgery (United States)*. SAGE Publications Inc., 151(2), pp. 348–353. doi: 10.1177/0194599814531912.

Blyth, M. (2019) *Auditory fitness for duty: Acoustic stealth awareness*, *University of Southampton*. doi: 10.1016/j.jsv.2010.04.020.

Bonate, P. L. (2001) 'A brief introduction to Monte Carlo simulation', *Clinical Pharmacokinetics*, 40(1), pp. 15–22. doi: 10.2165/00003088-200140010-00002.

Van Brink, J. and Gisselsson-Solen, M. (2019) 'Quality of life in Swedish children receiving grommets – An analysis of pre- and postoperative results based on a national quality register', *International Journal of Pediatric Otorhinolaryngology*. Elsevier, 120(February), pp. 44–50. doi: 10.1016/j.ijporl.2019.02.009.

British Society of Audiology (2018) 'Practice Guidance: Assessment Guidelines for the Distraction Test of Hearing', 86(August). Available at: www.thebsa.org.uk.

Brouwer, Carole N.M., Maillé, A. R., Rovers, M. M., Veenhoven, R. H., *et al.* (2005) 'Effect of pneumococcal vaccination on quality of life in children with recurrent acute otitis media: A randomized, controlled trial', *Pediatrics*, 115(2), pp. 273–279. doi: 10.1542/peds.2004-0778.

Brouwer, Carole N.M., Maillé, A. R., Rovers, M. M., Grobbee, D. E., *et al.* (2005) 'Health-related quality of life in children with otitis media', *International Journal of Pediatric Otorhinolaryngology*, 69(8), pp. 1031–1041. doi: 10.1016/j.ijporl.2005.03.013.

Brouwer, C.N.M. *et al.* (2005) 'The impact of recurrent acute otitis media on the quality of life of children and their caregivers', *Clinical Otolaryngology*, 30(3), pp. 258–265. doi: 10.1111/j.1365-2273.2005.00995.

Brouwer, C. N. M. *et al.* (2007) 'Reliability and validity of functional health status and health-related quality of life questionnaires in children with recurrent acute otitis media', *Quality of Life Research*, 16(8), pp. 1357–1373. doi: 10.1007/s11136-007-9242-0.

BSA (2016) 'Recommended procedure ear examination', *British Society of Audiology*, (January), pp. 1–12. Available at: http://www.thebsa.org.uk/wp-content/uploads/2014/04/RecProc_EarExam_25Jan2010.pdf.

BSA (2018) 'Pure-tone air-conduction and bone- conduction threshold audiometry with and without masking', *British Society of Audiology*, 39(August), pp. 1–39.

Cai, T. *et al.* (2017) 'Hearing Loss in Children With Otitis Media With Effusion: Actual and Simulated Effects on Speech Perception', *Ear and Hearing*, 39(4), pp. 645–655. doi: 10.1097/AUD.0000000000000519.

Cai, T. *et al.* (2018) 'Pure tone hearing profiles in children with otitis media with effusion', *Disability and Rehabilitation*. Informa UK Ltd., 40(10), pp. 1166–1175. doi: 10.1080/09638288.2017.1290698.

Cai, T. and McPherson, B. (2017) 'Hearing loss in children with otitis media with effusion: a systematic review', *International Journal of Audiology*, 56(2), pp. 65–76. doi: 10.1080/14992027.2016.1250960.

Carhart, R. (1951) 'Basic principles of speech audiometry', *Acta Oto-Laryngologica*, 40(1–2), pp. 62–71. doi: 10.3109/00016485109138908.

Carhart, R. and Tillman, T. W. (1970) 'Interaction of Competing Speech With Hearing Losses', *Arch Otolaryngol Head Neck Surg*, 91.

Chessman, R. *et al.* (2016) 'Which outcome measures are reported by clinical trials investigating OME treatment? A case for standardised reporting', *International Journal of Pediatric Otorhinolaryngology*, 86, pp. 93–96. doi: 10.1016/j.ijporl.2016.04.029.

Chow, A. H. C. *et al.* (2019) 'Otitis media with effusion in children: Cross-frequency correlation in pure tone audiometry', *PLoS ONE*, 14(8), pp. 1–13. doi: 10.1371/journal.pone.0221405.

Chow, Y., Wabnitz, D. A. M. and Ling, J. (2007) 'Quality of life outcomes after ventilating tube insertion for otitis media in an Australian population', *International Journal of Pediatric Otorhinolaryngology*, 71(10), pp. 1543–1547. doi: 10.1016/j.ijporl.2007.06.001.

Cienkowski, K. M., Ross, M. and Lerman, J. (2009) 'The Word Intelligibility by Picture Identification (WIPI) Test Revisited', *Journal of Educational Audiology*, 15, pp. 39–43.

Clark, J. G. (1981) *Uses and abuses of hearing loss classification*, *Asha: a journal of the American Speech-Language-Hearing Association*. Available at: https://www.researchgate.net/publication/16145943 (Accessed: 14 December 2018).

Crawford, B. *et al.* (2017) 'Impact of Pediatric Acute Otitis Media on Child and Parental Quality of Life and Associated Productivity Loss in Malaysia: A Prospective Observational Study', *Drugs - Real World Outcomes*, 4(1), pp. 21–31. doi: 10.1007/s40801-016-0099-9.

Cullington, H. E. *et al.* (2017) 'United Kingdom national paediatric bilateral project: Demographics and results of localization and speech perception testing', *Cochlear Implants International*, 18(1), pp. 2–22. doi: 10.1080/14670100.2016.1265055.

Dietz, A. *et al.* (2014) 'The development and evaluation of the Finnish Matrix Sentence Test for speech intelligibility assessment', *Acta Oto-Laryngologica*, 134(7), pp. 728–737. doi: 10.3109/00016489.2014.898185.

Dingemanse, G. and Goedegebure, A. (2020) 'Efficient Adaptive Speech Reception Threshold Measurements Using Stochastic Approximation Algorithms', *Trends in Hearing*, 23. doi: 10.1177/2331216520919199.

Van Dongen, T. M. A. *et al.* (2014) 'A trial of treatment for acute otorrhea in children with tympanostomy tubes', *New England Journal of Medicine*, 370(8), pp. 723–733. doi: 10.1056/NEJMoa1301630.

Eapen, R. J. *et al.* (2008) 'The development of frequency weighting for speech in children with a history of otis media with effusion (Ear and Hearing (2008) 29, (718-724))', *Ear and Hearing*, 29(6), p. 979. doi: 10.1097/01.aud.0000339411.98443.eb.

Els, T. and Olwoch, I. P. (2018) 'The prevalence and impact of otitis media with effusion in children admitted for adeno-tonsillectomy at Dr George Mukhari Academic Hospital, Pretoria, South Africa', *International Journal of Pediatric Otorhinolaryngology*. Elsevier, 110(February), pp. 76–80. doi: 10.1016/j.ijporl.2018.04.030.

Fabus, R. *et al.* (2018) 'The effectiveness of parental questionnaires in the assessment of speech-language and auditory function in children', *Folia Phoniatrica et Logopaedica*, 69(5–6), pp. 261–

270. doi: 10.1159/000488054.

Farinetti, A. *et al.* (2018) 'International consensus (ICON) on audiological assessment of hearing loss in children', *European Annals of Otorhinolaryngology, Head and Neck Diseases*, 135(1), pp. S41–S48. doi: 10.1016/j.anorl.2017.12.008.

Faul, F. *et al.* (2009) 'Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses', *Behavior Research Methods*, 41(4), pp. 1149–1160. doi: 10.3758/BRM.41.4.1149.

Field, A. (2009) *Discovering Stistics using SPSS*. Third Edit. London: SAGE Publications Inc. doi: 10.1007/978-0-387-68969-2_13.

Flaherty, M. M., Nudelman, C. and Bottalico, P. (2022) 'The effect of the number of items on the reliability of speech intelligibility results for WIPI test', *The Journal of the Acoustical Society of America*, 152(4), pp. A175–A175. doi: 10.1121/10.0015942.

Francis, N. A. *et al.* (2018) 'Oral steroids for hearing loss associated with otitis media with effusion in children aged 2–8 years: The OSTRICH RCT', *Health Technology Assessment*, 22(61), pp. 1–113. doi: 10.3310/hta22610.

Gan, R. W. C. *et al.* (2018) 'Quality of questionnaires for the assessment of otitis media with effusion in children', *Clinical Otolaryngology*, 43(2), pp. 572–583. doi: 10.1111/coa.13026.

García-Pérez, M. A. (1998) 'Forced-choice staircases with fixed step sizes: Asymptotic and small-sample properties', *Vision Research*, 38(12), pp. 1861–1881. doi: 10.1016/S0042-6989(97)00340-4.

Gelfand, A. (2009) *Hearing: An Introduction to Psychological and Physiological Acoustics*. 5th edn, *Taylor & Francis*. 5th edn. NewYork. doi: 10.1136/jnnp.45.12.1175-b.

Gelfand, S. (2016) 'Speech audiometry', in *Essentials of Audiology*. 4th edn. NewYork: Thieme Medical Publishers, pp. 215–247. doi: 10.4295/audiology.52.563.

Gelman, A. and Hill, J. (2007) 'Why?', in *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press Textbooks, pp. 1–11. doi: 10.53689/pys.v26i1.12.

Gescheider, G. A. (1997) *Psychoacoustics: The Fundamentals*. 3rd edn, *Taylor & Francis*. 3rd edn. Florence.

Goldsworthy, R. L. and Markle, K. L. (2019) 'Pediatric hearing loss and speech recognition in quiet and in different types of background noise', *Journal of Speech, Language, and Hearing Research*, 62(3), pp. 758–767. doi: 10.1044/2018_JSLHR-H-17-0389.

Haggard, M. (2004) 'Speech reception in noise: An indicator of benefit from otitis media with effusion surgery', *Clinical Otolaryngology and Allied Sciences*, 29(5), pp. 497–504. doi: 10.1111/j.1365-2273.2004.00843.x.

Hall, Amanda J. *et al.* (2014) 'Glue ear, hearing loss and IQ: An association moderated by the child's home environment', *PLoS ONE*, 9(2). doi: 10.1371/journal.pone.0087021.

Hall, A. J., Munro, K. J. and Heron, J. (2007) 'Developmental changes in word recognition threshold from two to five years of age in children with different middle ear status', *International Journal of Audiology*, 46, pp. 355–361. doi: 10.1080/14992020701331570.

Hall, D. A. *et al.* (2018) 'A good practice guide for translating and adapting hearing-related questionnaires for different languages and cultures', *International Journal of Audiology*, 57(3), pp. 161–175. doi: 10.1080/14992027.2017.1393565.

Harlor, A. D. *et al.* (2009) 'Clinical report - Hearing assessment in infants and children: Recommendations beyond neonatal screening', *Pediatrics*, 124(4), pp. 1252–1263. doi: 10.1542/peds.2009-1997.

Harries, J. and Williamson, T. (2000) 'Community-based validation of the McCormick toy test', *British Journal of Audiology*, 34(5), pp. 279–283. doi: 10.3109/03005364000000139.

Heidemann, C.H. *et al.* (2013) 'The Otitis Media-6 questionnaire: Psychometric properties with emphasis on factor structure and interpretability', *Health and Quality of Life Outcomes*, 11(1), pp. 1–10. doi: 10.1186/1477-7525-11-201.

Hochmuth, S. *et al.* (2012) 'A Spanish matrix sentence test for assessing speech reception thresholds in noise', *International Journal of Audiology*, 51(7), pp. 536–544. doi: 10.3109/14992027.2012.670731.

Holland Brown, T. *et al.* (2019) 'Using a Bone-Conduction Headset to Improve Speech Discrimination in Children With Otitis Media With Effusion', *Trends in Hearing*, 23, pp. 1–9. doi: 10.1177/2331216519858303.

Holland Brown, T., Marriage, J. and Salorio-Corbetto, M. (2022) 'Speech discrimination and word identification with a consumer-level bone-conduction headset and remote microphone for children with normal hearing', *International Journal of Audiology*. Taylor & Francis, 0(0), pp. 1–8. doi: 10.1080/14992027.2022.2049379.

Homøe, P. *et al.* (2020) 'Impact of Otitis Media on Quality of Life and Development', *Int J Pediatr Otorhinolaryngol*, 130, pp. 139–148. doi: 10.1016/j.ijporl.2019.109837.Panel.

Houston, D. M. *et al.* (2016) 'Factors affecting speech discrimination in children with cochlear implants: Evidence from early-implanted infants', *Journal of the American Academy of Audiology*, 27(6), pp. 480–488. doi: 10.3766/jaaa.15088.

Humaid, A.-H. I. *et al.* (2014) 'Prevalence and risk factors of Otitis Media with effusion in school children in Qassim Region of Saudi Arabia.', *International journal of health sciences*. Qassim University, 8(4), pp. 325–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25780352 (Accessed: 1 June 2018).

Indius, J. H. *et al.* (2018) 'Middle ear disease in Danish toddlers attending nursery day-care - Applicability of OM-6, disease specific quality of life and predictors for middle ear symptoms.', *International journal of pediatric otorhinolaryngology*. Elsevier, 110(April), pp. 130–134. doi: 10.1016/j.ijporl.2018.04.031.

Jabbari Moghaddam, Y. and Mirghaffari, A. (2018) 'Evaluation of Children Quality of Life after Serous Otitis Media Surgery', *Journal of Caring Sciences*, 7(3), pp. 131–135. doi: 10.15171/jcs.2018.021.

Jamal, A., Alsabea, A. and Tarakmeh, M. (2022) 'Effect of Ear Infections on Hearing Ability: A Narrative Review on the Complications of Otitis Media', *Cureus*, 14(7). doi: 10.7759/cureus.27400.

Katz, J. *et al.* (2009) *Handbook of Clinical Audiology*. 6th edn. Philadelphia.

Khavarghazalani, B. *et al.* (2022) 'Auditory temporal processinzg in children with history of recurrent otitis media with effusion', *Hearing, Balance and Communication*. Taylor & Francis, 20(1), pp. 46–51. doi: 10.1080/21695717.2022.2029091.

Kim, J. M. *et al.* (2016) 'The Best-Matched Pure Tone Average and Speech Recognition Threshold for Different Audiometric Configurations', *Korean Journal of Otorhinolaryngology-Head and Neck Surgery*, 59(10), p. 725. doi: 10.3342/kjorl-hns.2016.59.10.725.

Kingdom, A. A. F. and Prins, N. (2010) *Psychophysics A Practical Introduction*. second, *Managing*. second. Edited by M. Tucker. Mica Haley.

Kishon-Rabin, L. and Rosenhouse, J. (2000) 'Speech perception test for Arabic-speaking children.', *Audiology : official organ of the International Society of Audiology*, 39(5), pp. 269–77. doi: 10.3109/00206090009073091.

Klausen, O. *et al.* (2000) 'Lasting effects of otitis media with effusion on language skills and listening performance', *Acta Oto-Laryngologica*, 120(543), pp. 73–76. doi: 10.1080/000164800454026.

Koiek, S. *et al.* (2022) 'Long-term hearing abilities in noise in school-age children with early-childhood otitis media', *Journal of Hearing Science*.

Kollmeier, B. *et al.* (2016) 'Sentence Recognition Prediction for Hearing-impaired Listeners in Stationary and Fluctuation Noise with FADE', *Trends in Hearing*, 20, pp. 1–17. doi: 10.1177/2331216516655795.

Koo, T. K. and Li, M. y. (2016) 'A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research', *Journal of Chiropractic Medicine*. Elsevier B.V., 15(2), pp. 155–163. doi: 10.1016/j.jcm.2016.02.012.

Kubba, H., Swan, I. R. C. and Gatehouse, S. (2004) 'How appropriate is the OM6 as a discriminative instrument in children with otitis media?', *Archives of Otolaryngology - Head and Neck Surgery*. doi: 10.1001/archotol.130.6.705.

Kujala, T. *et al.* (2014) 'Quality of life after surgery for recurrent otitis media in a randomized controlled trial', *Pediatric Infectious Disease Journal*, 33(7), pp. 715–719. doi: 10.1097/INF.0000000000000265.

Kwon, M. and Yang, W. (2023) 'Effects of face masks and acoustical environments on speech recognition by preschool children in an auralised classroom', *Applied Acoustics*. Elsevier Ltd, 202, p. 109149. doi: 10.1016/j.apacoust.2022.109149.

Lameiras, A. R. *et al.* (2017) 'Validation of the Otitis media-6 questionnaire for European Portuguese | Validação do questionário Otitis media-6 para Português Europeu', *Acta Medica Portuguesa*, 30(5), pp. 381–387. doi: 10.20344/amp.7921.

Lameiras, A. R. *et al.* (2018) 'Quality of Life of Children with Otitis Media and Impact of Insertion of Transtympanic Ventilation Tubes in a Portuguese Population', *Notes and Queries*, 31(1), pp. 30–37. doi: 10.1093/nq/s5-VII.164.138-a.

Lauritsen, M.-B. G. *et al.* (2016) 'The Galker test of speech reception in noise; associations with background variables, middle ear status, hearing, and language in Danish preschool children', *International Journal of Pediatric Otorhinolaryngology*, 80, pp. 53–60. doi: 10.1016/j.ijporl.2015.11.014.

Leek, M. R. (2001) 'Adaptive procedures in psychophysical research', *Perception and*

Psychophysics, 63(8), pp. 1279–1292. doi: 10.3758/BF03194543.

Leek, M. R., Thomas, H. E. and Lynne, M. (1991) 'An interleaved tracking procedure to monitor unstable psychometric functions', *Journal of the Acoustical Society of America*, 90(3), pp. 1385–1397. doi: 10.1121/1.401930.

Levitt, H. (1971) 'Transformed Up-Down Methods in Psychoacoustics', *The Journal of the Acoustical Society of America*, 49(2B), pp. 467–477. doi: 10.1121/1.1912375.

Liljequist, D., Elfving, B. and Roaldsen, K. S. (2019) *Intraclass correlation – A discussion and demonstration of basic features*, *PLoS ONE*. doi: 10.1371/journal.pone.0219854.

Liu, P. Z. *et al.* (2020) 'A core outcome set for research on the management of otitis media with effusion in otherwise-healthy children', *International Journal of Pediatric Otorhinolaryngology*. Elsevier, 134(March), p. 110029. doi: 10.1016/j.ijporl.2020.110029.

Loven, F. C. and Hawkins, D. B. (1983) 'Interlist equivalency of the CID W-22 word lists presented in quiet and in noise', *Ear and Hearing*, 4(2), pp. 91–97. doi: 10.1097/00003446-198303000-00005.

Lovett, R., Summerfield, Q. and Vickers, D. (2013) 'Test-retest reliability of the Toy Discrimination Test with a masker of noise or babble in children with hearing impairment', *International Journal of Audiology*, 52(6), pp. 377–384. doi: 10.3109/14992027.2013.769064.

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. doi: 10.1177/2331216514537722.

Maile, E. J. and Youngs, R. (2013) 'Quality of life measures in otitis media', *Journal of Laryngology and Otology*, 127(5), pp. 442–447. doi: 10.1017/S0022215113000509.

Martines, F. *et al.* (2011) 'Otitis media with effusion with or without atopy: audiological findings on primary schoolchildren', *American Journal of Otolaryngology*. W.B. Saunders, 32(6), pp. 601–606. doi: 10.1016/J.AMJOTO.2010.08.002.

Maxwell, S. E. and Delaney, H. D. (2004) *Designing experiments and analyzing data: A model comparison perspective, 2nd ed.*, *Designing experiments and analyzing data: A model comparison perspective, 2nd ed.* Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

McCormick, B. (1977) 'The toy discrimination test: An aid for screening the hearing of children above a mental age of two years', *Public Health*. W.B. Saunders, 91(2), pp. 67–74. doi:

10.1016/S0033-3506(77)80003-6.

McCoul, E. D. *et al.* (2011) 'A prospective study of the effect of gastroesophageal reflux disease treatment on children with otitis media', *Archives of Otolaryngology - Head and Neck Surgery*, 137(1), pp. 35–41. doi: 10.1001/archoto.2010.222.

Mendel, L. L. (2008) 'Current considerations in pediatric speech audiometry', *International Journal of Audiology*, 47(9), pp. 546–553. doi: 10.1080/14992020802252261.

Meng, X. L., Rosenthal, R. and Rubin, D. B. (1992) 'Comparing correlated correlation coefficients', *Psychological Bulletin*, 111(1), pp. 172–175. doi: 10.1037/0033-2909.111.1.172.

MOH (2020) *Coronavirus Disease (COVID-19)*, *Ministry of Health, Saudi Arabia*. Available at: https://covid19awareness.sa/wp-content/uploads/2020/06/What-is-Corona-English_New.pdf.

Moore, B. C. J. and Glasberg, B. R. (1993) 'Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech', *Citation: The Journal of the Acoustical Society of America*, 94, p. 1229. doi: 10.1121/1.407478.

Muhaimeid, H., Zakzouk, S. and Bafaqeeh, S. (1993) 'Epidemiology of chronic suppurative otitis media in Saudi children.', *International journal of pediatric otorhinolaryngology*. Elsevier, 26(2), pp. 101–8. doi: 10.1016/0165-5876(93)90015-U.

Neumann, K. *et al.* (2012) 'Speech audiometry in quiet with the Oldenburg Sentence Test for Children', *International Journal of Audiology*, 51(3), pp. 157–163. doi: 10.3109/14992027.2011.633935.

Newton, V. E. (2009) *Paediatric Audiological Medicine*. 2nd edn. Chichester, West Sussex: John Wiley & Sons. doi: 10.1016/0148-9062(96)89910-4.

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.

NICE (2018) '2018 surveillance of otitis media with effusion in under 12s: surgery (NICE guideline CG60) Surveillance report', (December). Available at: https://www.nice.org.uk/guidance/cg60/resources/2018-surveillance-of-otitis-media-with-effusion-in-under-12s-surgery-nice-guideline-cg60-6604581853/chapter/Surveillance-decision?tab=evidence&utm_source=Surveillance+report+alerts&utm_campaign=f9a5dbb682-EM.

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.

Ondáš, S. *et al.* (2020) 'Pediatric speech audiometry web application for hearing detection in the home environment', *Electronics (Switzerland)*, 9(6), pp. 1–15. doi: 10.3390/electronics9060994.

Ozimek, E., Warzybok, A. and Kutzner, D. (2010) 'Polish sentence matrix test for speech intelligibility measurement in noise', *International Journal of Audiology*, 49(6), pp. 444–454. doi: 10.3109/14992021003681030.

De Paepe, J. *et al.* (2019) 'Ear- and hearing-related impact on quality of life in children with cleft palate: Development and pretest of a health-related quality of life (HRQOL) instrument', *International Journal of Pediatric Otorhinolaryngology*. Elsevier, 122(January), pp. 35–39. doi: 10.1016/j.ijporl.2019.03.023.

Palmer, A. R., Sheppard, S. and Marshal, D. H. (1991) 'Short paper: Prediction of hearing thresholds in children using an automated toy discrimination test', *British Journal of Audiology*. doi: 10.3109/03005369109076609.

Pantaleon, L. (2019) 'Why measuring outcomes is important in health care', *Journal of Veterinary Internal Medicine*, 33(2), pp. 356–362. doi: 10.1111/jvim.15458.

Pekkarinen, E., Salmivalli, A. and Suonpää, J. (1990) 'Effect of noise on word discrimination by subjects with impaired hearing, compared with those with normal hearing', *Scandinavian Audiology*, 19(1), pp. 31–36. doi: 10.3109/01050399009070749.

Pelton, S. (2022) 'Acute otitis media in children: Epidemiology, microbiology, and complications', *UpToDate*. Available at: https://www.uptodate.com/contents/acute-otitis-media-in-children-epidemiology-microbiology-clinical-manifestations-and-complications?search=acute otitis media children&source=search_result&selectedTitle=3~150&usage_type=default&display_rank=3 (Accessed: 9 August 2018).

Penn, T. O., Grantham, D. W. and Gravel, J. S. (2004) 'Simulated conductive hearing loss in children.', *Journal of the American Academy of Audiology*, 15(4), pp. 300–305. doi: 10.3766/jaaa.15.4.4.

Perdrizet, J. *et al.* (2022) 'The broader impacts of otitis media and sequelae for informing economic evaluations of pneumococcal conjugate vaccines', *Expert Review of Vaccines*. Taylor & Francis, 21(4), pp. 499–511. doi: 10.1080/14760584.2022.2040989.

Perigoe, C. B. and Paterson, M. M. (2013) 'Understanding Auditory Development and the Child with Hearing Loss', *Fundamentals of audiology for the speech-language pathologist*, pp. 173–204.

Plomp, R. (1978) 'Auditory handicap of hearing impairment and the limited benefit of hearing aids', *The Journal of the Acoustical Society of America*, 63(2), pp. 533–549. doi: 10.1121/1.381753.

Plomp, R. and Mimpen, A. M. (1979) 'Improving the reliability of testing the speech reception threshold for sentences', *International Journal of Audiology*, 18(1), pp. 43–52. doi: 10.3109/00206097909072618.

Ravicz, M. E., Rosowski, J. J. and Merchant, S. N. (2004) 'Mechanisms of hearing loss resulting from middle-ear fluid', *Hearing Research*, 195(1–2), pp. 103–130. doi: 10.1016/j.heares.2004.05.010.

Raychaudhuri, S. (2008) 'Introduction to monte carlo simulation', *Proceedings - Winter Simulation Conference*, pp. 91–100. doi: 10.1109/WSC.2008.4736059.

Rayes, H., Al-malky, G. and Vickers, D. (2021) 'The Development of a Paediatric Phoneme Discrimination Test for Arabic Phonemic Contrasts', pp. 150–166.

Richards, M. and Giannoni, C. (2002) 'Quality-of-life outcomes after surgical intervention for otitis media', *Archives of Otolaryngology - Head and Neck Surgery*, 128(7), pp. 776–782. doi: 10.1001/archotol.128.7.776.

Ristovska, L. *et al.* (2021) 'Correlation between pure tone thresholds and speech thresholds', *Human Research in Rehabilitation*, 11(2), pp. 120–125. doi: 10.21554/hrr.092108.

Rosarioa, D. C. and Mendez, M. D. (2023) 'Chronic suppurative otitis media', *BMJ clinical evidence*. doi: 10.29309/tpmj/2010.17.03.2531.

Rosenfeld, R. M. *et al.* (2000) 'Impact of tympanostomy tubes on child quality of life', *Archives of Otolaryngology - Head and Neck Surgery*. doi: 10.1001/archotol.126.5.585.

Rosenfeld, R. M. *et al.* (2016) 'Clinical practice guideline', *Otolaryngology - Head and Neck Surgery (United States)*, 154(2), pp. 201–214. doi: 10.1177/0194599815624407.

Rosenfeld, R. M. *et al.* (2022) 'Clinical Practice Guideline: Tympanostomy Tubes in Children (Update)', *Otolaryngology - Head and Neck Surgery (United States)*, 166(1_suppl), pp. S1–S55. doi: 10.1177/01945998211065662.

Rosenfeld, R. M., Goldsmith, A. and Balzano, L. T. (1997) 'Quality of life for children with otitis media', *Archives of Otolaryngology - Head and Neck Surgery*, 123, pp. 1049–1054. doi:

10.1001/archotol.1997.01900100019002.

Rosenfeld, R. M., Goldsmith, A. J. and Madell, J. R. (1998) 'How accurate is parent rating of hearing for children with otitis media?', *Archives of Otolaryngology - Head and Neck Surgery*, 124(9), pp. 989–992. doi: 10.1001/archotol.124.9.989.

Rosenfeld, R. M. and Kay, D. (2003) 'Natural history of untreated otitis media', *Laryngoscope*, 113(10), pp. 1645–1657. doi: 10.1097/00005537-200310000-00004.

Rosenfeld, R. M., Mandel, E. M. and Bluestone, C. D. (1991) 'Systemic Steroids for Otitis Media With Effusion in Children', *Arch Otolaryngol Head Neck Surg*, 117, pp. 984–989.

Rovers, M. M. *et al.* (2001) 'Randomised controlled trial of the effect of ventilation tubes (grommets) on quality of life at age 1-2 years', *Archives of Disease in Childhood*, 84(1), pp. 45–49. doi: 10.1136/adc.84.1.45.

Ruotsala, A. (2016) *Digital Close-Range Photogrammetry – A Modern Method to Document Forensic Mass Graves*. University of Helsinki.

Saber, A. *et al.* (2021) 'A Novel Deep-Learning Model for Automatic Detection and Classification of Breast Cancer Using the Transfer-Learning Technique', *IEEE Access*, 9, pp. 71194–71209. doi: 10.1109/ACCESS.2021.3079204.

Sabo, D. L. *et al.* (2003) 'Hearing levels in infants and young children in relation to testing technique, age group, and the presence or absence of middle-ear effusion', *Ear and Hearing*, 24(1), pp. 38–47. doi: 10.1097/01.AUD.0000051988.23117.91.

Saraf, A., Manhas, M. and Kalsotra, P. (2022) 'Assessment of Quality of Life After Ventilation Tube Insertion Using Otitis Media 6-Item (OM-6) Questionnaire', *Indian Journal of Otolaryngology and Head and Neck Surgery*. doi: 10.1007/s12070-022-03278-9.

Savithri, S. R. (2022) 'Development of Speech Perception', *Research Advances in Communication Studies*, pp. 249–340. doi: 10.1007/978-3-030-81542-4_8.

Al Sayah, F. *et al.* (2013) 'Health related quality of life measures in Arabic speaking populations: A systematic review on cross-cultural adaptation and measurement properties', *Quality of Life Research*, pp. 213–229. doi: 10.1007/s11136-012-0129-3.

Schneider, A., Hommel, G. and Blettner, M. (2010) 'Lineare regressionsanalyse - Teil 14 der serie zur bewertung wissenschaftlicher publikationen', *Deutsches Arzteblatt*, 107(44), pp. 776–782. doi: 10.3238/arztebl.2010.0776.

Semeraro, H. D. (2015) *Developing a measure of auditory fitness for duty for military personnel*. Available at: https://eprints.soton.ac.uk/388043/1/HDSemeraro_PhD_ISVR_FEE_080216.pdf (Accessed: 20 July 2018).

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.

Shehabi, A. (2017) *Developing a Speech in Noise Test in Arabic*.

Shen, Y. (2013) 'Comparing adaptive procedures for estimating the psychometric function for an auditory gap detection task', *Attention, Perception, and Psychophysics*, 75(4), pp. 771–780. doi: 10.3758/s13414-013-0438-9.

Singleton, A. J. and Waltzman, S. B. (2015) 'Audiometric Evaluation of Children with Hearing Loss', *Otolaryngologic Clinics of North America*. Elsevier Inc, 48(6), pp. 891–901. doi: 10.1016/j.otc.2015.06.002.

Slavin, R. and Smith, D. (2009) 'The relationship between sample sizes and effect sizes in systematic reviews in education', *Educational Evaluation and Policy Analysis*, 31(4), pp. 500–506. doi: 10.3102/0162373709352369.

Smoorenburg, G. F. (1992) 'Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their tone audiogram', *Journal of the Acoustical Society of America*, 91(1), pp. 421–437. doi: 10.1121/1.402729.

Snapp, H., Vogt, K. and Agterberg, M. J. H. (2020) 'Bilateral bone conduction stimulation provides reliable binaural cues for localization', *Hearing Research*. Elsevier B.V, 388, p. 107881. doi: 10.1016/j.heares.2019.107881.

Strasburger, H. (2001) 'Converting between measures of slope of the psychometric function', *Perception and Psychophysics*, 63(8), pp. 1348–1355. Available at: www.grp.hwz.uni- (Accessed: 2 January 2019).

Streiner, D. L., Norman, G. and Cairney, J. (2015) *Health Measurement Scales, 5th Edition*, *Oxford University Press*. doi: 10.1136/jech.44.4.328.

Summerfield, Q. *et al.* (1994) 'Clinical evaluation and test-retest reliability of the IHR-McCormick Automated Toy Discrimination Test', *British Journal of Audiology*, 28, pp. 165–179. doi: 10.3109/03005369409086564.

Tao, J., Schulz, K., Jeffe, Donna B, *et al.* (2018) 'Validations of the OM-6 Parent-Proxy Survey for Infants/Toddlers with Otitis Media', *Original Research-Pediatric Otolaryngology Otolaryngology-Head and Neck Surgery*, 158(5), pp. 934–941. doi: 10.1177/0194599817750372.

Timmerman, A. A. *et al.* (2007) 'Psychometric qualities of questionnaires for the assessment of otitis media impact', *Clinical Otolaryngology*, 32(6), pp. 429–439. doi: 10.1111/j.1749-4486.2007.01570.x.

Timmerman, A. A. (2008) *How to describe the functional health status of children with otitis media*. Available at: http://dissertaties.ub.unimaas.nl/stelling/12866.pdf.

TO, P., DW, G. and JS, G. (2004) 'Simulated conductive hearing loss in children.', *Journal of the American Academy of Audiology*, 15(4), pp. 300–310. doi: 10.3766/jaaa.15.4.4.

Tronstad, T. V. (2017) 'Statistical tool to detect small hearing threshold shifts', *International Journal of Audiology*, 56(8), pp. 596–606. doi: 10.1080/14992027.2017.1303203.

Twisk, J. W. R. (2019) *Applied Mixed Model Analysis*, *Applied Mixed Model Analysis*. doi: 10.1017/9781108635660.

Venekamp, R. *et al.* (2016) 'Interventions for children with ear discharge occurring at least two weeks following grommet (ventilation tube) insertion (Review)', *Cochrane Database of Systematic Reviews*, (11). doi: 10.1002/14651858.CD011684.pub2.

Venekamp, R. *et al.* (2018) 'Grommets (ventilation tubes) for recurrent acute otitis media in children.', *The Cochrane database of systematic reviews*, (5). doi: 10.1002/14651858.CD012017.pub2.

Vickers, D. A. *et al.* (2017) 'Closed-Set Speech Discrimination Tests for Assessing Young Children', *Ear and Hearing*, 39(1), pp. 32–41. doi: 10.1097/AUD.0000000000000528.

Vlastos, I. M. *et al.* (2009) 'Quality of life in children with chronic suppurative otitis media with or without cholesteatoma', *International Journal of Pediatric Otorhinolaryngology*, 73, pp. 363–369. doi: 10.1016/j.ijporl.2008.10.030.

Vlastos, I. M. *et al.* (2011) 'Adenoidectomy plus tympanostomy tube insertion versus adenoidectomy plus myringotomy in children with obstructive sleep apnoea syndrome', *Journal of Laryngology and Otology*, 125(3), pp. 274–278. doi: 10.1017/S0022215110002549.

Wagener, K., Josvassen, J. L. and Ardenkjar, R. (2003) 'Design, optimization and evaluation of a Danish sentence test in noise', *International Journal of Audiology*, 42, pp. 10–17.

Watkins, G. D., Swanson, B. A. and Suaning, G. J. (2016) 'A simulation analysis of the variability of the roving level hearing test', *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*. IEEE, 2016-Octob, pp. 4715–4718. doi: 10.1109/EMBC.2016.7591780.

Watkins, G. D., Swanson, B. A. and Suaning, G. J. (2020) 'Prediction of Individual Cochlear Implant Recipient Speech Perception with the Output Signal to Noise Ratio Metric', *Ear and Hearing*, 41(5), pp. 1270–1281. doi: 10.1097/AUD.0000000000000846.

WHO (2004) 'Chronic suppurative otitis media - Burden of Illness and Management Options', *WHO Library Cataloguing-in-Publication Data*, p. 84. doi: 10.1016/j.amjoto.2007.09.002.

Williamson, I. *et al.* (1997) 'A New Preformance-in-noise Test for assessing hearing disability in young schoolchildren in a community setting.pdf', *Otitis Media Today*, pp. 111–115.

Williamson, I. and Sheridan, C. (1994) 'The development of a test of speech reception disability for use in 5- to 8-year-old children with otitis media with effusion', *International Journal of Language & Communication Disorders*, 29(1), pp. 27–37. doi: 10.3109/13682829409041479.

Wilson, R. H. (2003) 'Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance', *Journal of the American Academy of Audiology*, 14(9), pp. 453–470. doi: 10.1055/s-0040-1715938.

Wilson, R. H. and Carter, A. S. (2001) 'Relation between slopes of word recognition psychometric functions and homogeneity of the stimulus materials', *Journal of the American Academy of Audiology*, 12(1), pp. 7–14. doi: 10.1055/s-0041-1741115.

Wong, L. L. N. and Soli, S. D. (2005) 'Development of the Cantonese Hearing In Noise Test (CHINT)', *Ear & Hearing*, 26(3), pp. 276–289. doi: 0196/0202/05/2603-0276/0.

Yazici, A. and Coskun, M. E. (2018) 'The effect of ventilation tube insertion to the health-related quality of life in a group of children in Southeast Anatolia', *Clinical Otolaryngology*. Blackwell Publishing Ltd, 43(6), pp. 1578–1582. doi: 10.1111/coa.13220.

Zakzouk, S. M. and Abduljawad, K. A. (2002) 'Point prevalence of type B tympanogram in children', *Saudi Medical Journal*, 23(6), pp. 708–710.

Zakzouk, S. M., Jamal, T. S. and Daghistani, K. J. (2002) 'Epidemiology of acute otitis media among Saudi children', *International Journal of Pediatric Otorhinolaryngology*, 62(3), pp. 219–222. doi: 10.1016/S0165-5876(01)00617-6.

Zaltz, Y. *et al.* (2019) 'Logarithmic versus linear change in step size when using an adaptive threshold-seeking procedure in a frequency discrimination task: Does it matter?', *Journal of Speech, Language, and Hearing Research*, 62(10), pp. 3887–3900. doi: 10.1044/2019_JSLHR-H-19-0049.