TOWARDS A PHYSIOLOGICAL MEASURE OF LISTENING EFFORT

Kelly Miles

Bachelor of Speech and Hearing Sciences (First Class Honours)

A thesis submitted in fulfilment of the requirements for the degree of:

DOCTOR OF PHILOSOPHY

Audiology Section, Department of Linguistics

Faculty of Human Sciences

Macquarie University

Sydney, Australia

Submitted: 14 July 2017

Abstract

This dissertation investigated the clinical feasibility of two physiological measures of listening effort; changes in pupil dilation and changes in alpha power (a cortical oscillation between 8-12 Hz). It was motivated by two factors: (1) sustained listening effort can lead to adverse health consequences or could adversely affect healthy behaviours (such as withdrawing socially) yet it is not clinically assessed; and (2) listening effort is a multi-faceted construct, and it remains unknown whether current physiological measures claimed to assess listening effort evaluate the same processes. We conducted laboratory studies on young adults with normal hearing and cognition to better understand how manipulating task difficulty can lead to changes in subjective and physiological measures of listening effort. Specifically, listener-internal factors were investigated by evaluating how working memory capacity, measured with a reading span task, interacted with subjective and physiological measures of listening effort. Listener-external factors were evaluated by examining the effects of channel-vocoding and performance parameters on subjective and physiological measures of listening effort. We also assessed how different data processing strategies and statistical approaches used across studies affect the results and interpretation of physiological outcome measures. The results of these studies indicated that while working memory capacity predicted perceived listening effort ratings, this was not the case for the two physiological measures. Perceived listening effort ratings appeared to be driven by estimated performance, and not the effort required to perform a speech recognition task. On the other hand, the physiological measures were both sensitive to changes in channel-vocoding, and the pupil response was further sensitive to performance levels and task accuracy. The physiological measures were not correlated with each other, suggesting each may be assessing a different aspect of listening effort. Finally, differences in data processing and statistical approach greatly altered the results and subsequent interpretation of the findings.

This dissertation provides an opportunity to advance the understanding of listening effort in an experimental setting, and was conducted within the overarching context of exploring the viability of a physiological tool to assess listening effort in a clinical environment.

Statement from Author

I state that this work has been submitted exclusively to Macquarie University (Sydney, Australia) for the consideration of a PhD degree.

Ethical review, guidance and approval have been obtained from Macquarie University Ethics Review Committee (Human Research). No. 5201100426.

I certify that I developed the original idea in collaboration with my three supervisors (A/Prof Catherine McMahon, Dr Isabelle Boisvert and Prof Björn Lyxell), and taken leadership to conduct all parts of this research work, including writing the content of this thesis. My supervisors have assisted in improving the research protocol, analyses, and interpretation of the data, as well as the quality of the written presentations. Co-authors (Dr Peter de Lissa, Dr Ronny Ibrahim, Dr Petra Graham, Dr Kenneth Beath, Mr Timothy Beechey and Mr Nay San) and reviewers of the papers (A/Prof Robert Cowan, and anonymous reviewers from the publications' journals) have helped in improving the manuscripts. I conducted the majority of the data collection, with support from Ms Louise Granger and Mr Chi Yun Lo. Statistical support has been obtained from Dr Kenneth Beath, Mr Timothy Beechey, Dr Örjan Dahlström, A/Prof Henrik Danielsson and Dr Petra Graham.

Kelly Miles

Table of Contents

Foreword and acknowledgements

First and foremost, I would like to thank my relentless cheerleader, Carly Miles, who has helped keep me motivated throughout my entire candidature. From late night flash cards during my undergraduate years, all the way up to submitting my doctoral thesis, you have always supported my academic adventures and been there for me every step of the way. I am eternally grateful. Thanks to my brother, mother, father and Margaret Collins for absolutely everything you have done to help me on this journey. I could never have imagined the extent of your love and support and could not have done this without you.

My supervisors, Catherine McMahon, Isabelle Boisvert, and Björn Lyxell, I am truly grateful to have benefitted from your combined experience and knowledge. I could not have imagined a better supervisory panel to guide my studies. Catherine – you are truly inspirational with your breadth of knowledge and dedication to evidence-based practice. I feel honoured that you took me on as a PhD student, and cannot thank you enough for the time you have invested in me. Isabelle – thank you for taking me on as your first PhD student. You have helped me grow and learn in so many ways. I know at times it has (I have) been difficult, but you never shied away from the challenges. I am a stronger researcher because of you. And to Björn – without your expertise, I would never have been able to piece all the puzzles of my thesis together. Tack för allt. Jag uppskattar verkligen dig!

I would like to thank all of the amazing people at MQ who have helped me on this journey: Ben Davies, Shiree Heath, Susan Lin, Caroline Moir, Collette Ryan, Tamara Schembri, Anita Szakay, Deanna Wong, Margaret Wood, Ivan Yuen and Nan Xu. Whether academically, administratively, or socially, you all contributed to making my PhD journey that much better. In particular, I'd like to thank Katherine Demuth who first hired me as a research assistant during my undergraduate years. You sparked my interest in a research career and I will be forever thankful. Felicity Cox – there will never be a better teacher than you. Your encouragement over the last 10(!) years has meant the world to me.

To the old and new members of the Audiology and Hearing Research Group, thank you for the dynamic group discussions and Tim Tams (and occasionally grapes). A very special thanks to Jaime Undurraga, Nay San, and Keiran Thompson, for throwing me into the deep end of R and not letting me drown. Thank you to Petra Graham and Kenneth Beath for inspiring statistical discussion. Ronny Ibrahim – an extra special thanks to you. Your impromptu crash courses in signals processing, MATLAB, and Sydney's best restaurants, were some of the best days during my candidature.

Thank you to the Linnaeus Centre HEAD, Swedish Institute for Disability Research, Linköping University group for making me feel most welcome on my visits. In particular, Mary Rudner for taking the time to discuss work (and the bushwalks), Henrik Danielsson and Örjan Dahlström for statistical help, and, of course, the Doktorandfika group. You all made my time abroad unforgettable.

To my friends who have been there throughout, and tolerated my absence when my PhD got in the way of socialising, I cannot express how much you all mean to me. Fadwa Alnafjan, Michaela Cooper, Leigh Fernandez, Victoria Heath, Kirsty Parra, and Trudy Krajenbrink. You are all amazingly inspirational women and push me to strive to the greatest of heights. I thank you.

Tim Beechey, Nick Haywood, Peter de Lissa, Matthieu Recugnat and Adam Weisser. You made Thursdays my favourite day of the week.

And to Sebastian. The thought of having done this without you is unconscionable. Thank you for keeping me sane.

List of Manuscripts

1. **Miles, K**., McMahon, C., Boisvert, I., & Lyxell, B. *Perceived listening effort ratings and the influence of working memory capacity.* Unpublished.
2. **Miles**, **K**., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., & Lyxell, B. (2017). *Objective assessment of listening effort: co-registration of pupillometry and EEG.* Published: Trends in Hearing.
3. **Miles, K**., Boisvert, I., McMahon, C., & Lyxell, B. *Physiological measures of listening effort and interactions with working memory capacity: a short communication.* Unpublished.
4. **Miles, K**., Boisvert, I., McMahon, C., San, N., Beath, K., Beechey, T., & Lyxell, B. *Pupillometry as a measure of listening effort: methodological considerations for data and statistical analysis.* In submission: Psychophysiology.

Additional Scientific Contributions

Publications

2016 McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., **Miles, K**., & Graham, P. L. Monitoring alpha oscillations and pupil dilation across the performance-intensity function. *Frontiers in Psychology*, 7: 745.

Conference presentation and posters

2017 **Miles, K**., Boisvert, I., McMahon, C., San, N., Beath, K., Beechey, T., & Lyxell, B. *Pupillometry as a measure of listening effort: methodological considerations for data and statistical analysis*. Fourth International Conference on Cognitive Hearing Science for Communication, Linköping, Sweden, 18-21 June.

2016 McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., **Miles, K**., Lo, C., & Granger, L. *Towards an objective measure of listening effort.* Audiology Australia National Conference, Melbourne, Australia, 22-24 May.

2016 **Miles, K.,** McMahon, C., Boisvert, I., & Lyxell, B. *Increasing assessment sensitivity using an objective measure of listening effort.* Audiology Australia National Conference, Melbourne, Australia, 22-24 May.

2015 **Miles, K.,** McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., & Lyxell, B. *Effect of adverse listening conditions and cognitive processes on pupil dilation and the alpha oscillatory network.* Third International Conference on Cognitive Hearing Science for Communication, Linköping, Sweden, 14-17 June.

2014 Boisvert, I., McMahon, C., de Lissa, P., **Miles, K.,** & Ibrahim, R. *Alpha oscillations and pupil dilation to measure listening effort; an initial study with adults using CIs*. 8th International Symposium on Objective Measures in Auditory Implants, Toronto, Canada, 15-18 October.

2014 McMahon, C., Ibrahim, R., Boisvert, I., & **Miles, K.** *Can we reliably measure listening effort during a clinical speech perception task?* XXXII World Congress of Audiology, Brisbane, Australia, 03-07 May.

2013 McMahon, C., Ibrahim, R., Peter, V., Al-meqbel, A., & **Miles**, **K.** *Objective measurement of temporal auditory processing in young adults*. Inaugural Australian Hearing Hub Conference, Sydney, Australia, 17-19 April.

Invited talks

2015 **Audiology Australia,** Sydney, Australia.
Title: Towards an objective measure of listening effort for individuals with hearing loss.

2015 **Linnaeus Centre for Hearing and Deafness**, Swedish Institute for Disability Research, Linköping, Sweden.
Title: Towards an objective measure of listening effort.

2015 **Örebro University**, Örebro, Sweden.
Title: Subjective and objective measures of listening effort: interactions with working memory capacity.

2014 **Sydney Cochlear Implant Centre**, Sydney, Australia.
Title: Objectively measuring listening effort during a standard clinical speech test.

List of Tables

List of Figures

Chapter 1

Chapter 3

Chapter 4

Chapter 6

List of Abbreviations

LME: Linear mixed-effects

P-LE: Pupillometry/ listening effort

SNR: Signal-to-noise ratio

SRT: Speech reception threshold

rmANOVA: Repeated measures analysis of variance

WMC: Working memory capacity

Chapter 1 Introduction

1.1 Preamble

"It's not the deafness that's the problem, it's the effort required to get anything from the hearing. It's all effort"[1]

Communication is central to the social world in which we live, and is critical along the life-course from infancy through to old age. Hearing impairment and poor acoustic environments can disrupt fluid communication by impeding our ability to hear and comprehend speech, ultimately affecting education, wealth development, independence, and social connectedness. While considerable progress has been made in managing hearing loss, many challenges remain. For example, individuals can differ in their ability to perceive and understand speech, even when they present with a similar degree and type of hearing loss and are fitted with a similar type of hearing device and prescription. The reasons for this are varied, and indeed complex, but may relate to the pathophysiology of the hearing loss itself (Edwards, 2007), the individual's cognitive function (Lunner, Rudner, & Rönnberg, 2009), or the interaction between hearing loss and cognitive ability (Stenfelt & Rönnberg, 2009), particularly in older populations where each is more prevalent (Uhlmann, Larson, Rees, Koepsell & Duckert, 1989; Pichora-Fuller, 2003).

While hearing sensitivity is quantifiable, the standard audiological test battery (typically encompassing pure tone air and bone conduction thresholds, immittance and monosyllabic words or short sentences in quiet or in noise) falls short in assessing the full effects of hearing loss on the individual and their communicative capacity. That is, current assessments do not: (1) provide comprehensive information about hearing ability (e.g., recent evidence in mice

[1] Patient quote from a focus group discussing perceived listening effort.
Hughes, S., Hutchings, H., & Rapport, F. (2016). Seeking connectedness: A constructivist Grounded Theory of perceived listening effort in cochlear implantation. Poster session presented at the British Society of Audiology Conference, Coventry, United Kingdom.

suggests that ageing causes a reduction in synaptic density which may reduce the dynamic range, potentially affecting one's ability to understand speech in a noisy environment, rather than affecting hearing thresholds (Sergeyenko, Lall, Liberman, & Kujawa, 2013)); (2) incorporate meaningful information about how cognition can influence speech perception, particularly in noise (e.g., using contextual cues to 'fill in the gaps'); nor (3) provide information about the daily experience of the individual, and how personal and environmental factors may influence communication (aligned with the World Health Organisation's International Classification of Functioning; (WHO, 2001)). While self-reported hearing handicap measures exist (e.g., Hearing Handicap Inventory for the Elderly; Ventry & Weinstein, 1982), these are based on retrospective recall of events and experiences, do not capture the daily variation of hearing ability/ handicap that occurs, and can be inaccurate particularly in older populations who tend to under-report the extent of their impairment (Uchida, Nakashima, Ando, Niino, & Shimokata, 2003).

The conceptualisation of listening effort as a key construct within the field of audiology has rapidly emerged to address some of the limitations of the standard audiological test battery, and facilitate personalisation of hearing devices and care for people with hearing loss (Danermark et al., 2010). Assessing listening effort in the clinic may verify a person's ease of listening during speech perception tasks, allowing both the clinician and individual to make better and more informed decisions regarding device selection and optimal device settings. Given the primary complaint of people with hearing impairment is listening to speech in noise (Dawes et al., 2014), and the adverse health effects of sustained effort including stress (Hua et al., 2014), cardiovascular strain (Peters et al., 1998), and fatigue (Hua, Anderzén-Carlsson, Widén, Möller, & Lyxell, 2015; Kramer, Kapteyn, & Houtgast, 2006; Pichora-Fuller, 2003), then increasing the sensitivity of current assessment measures may enhance conversational capacity and social engagement, leading to improved psychosocial and

physical outcomes. Given the rapid developments in this field over the time course of this thesis, a definition (McGarrigle et al., 2014) and a framework developed through expert consensus (Pichora-Fuller et al., 2016) for understanding/conceptualising listening effort have been proposed. At the current time, studies are being conducted to explore the robustness of the definitions, and certainly some debate has already arisen (see Rönnberg et al., in McGarrigle et al., 2014). Nonetheless, there is consensus that listening effort embodies a multitude of processes which are encompassed by auditory (bottom-up) and cognitive (top-down) mechanisms.

However, our understanding of what listening effort is, and therefore which measures can be used to evaluate it, are lacking. While several tools which have claimed to measure listening effort exist, the multifaceted nature of listening effort makes it particularly challenging to determine what each tool is specifically measuring. In addition, levels of arousal, motivation, and how and what cognitive and linguistic processes are employed to perform a demanding listening task may vary between individuals (Pichora-Fuller et al., 2016). Variability in outcome measures may therefore reflect different (or overlapping) cognitive processing strategies, variations in cognitive capacity, interactions with motivation and fatigue, and/or the insensitivity of the instrument itself. Further to this, the way in which outcome measures are processed and analysed also vary between studies and may contribute to disagreement regarding which cognitive processes each measure is assessing, and which variables modulate effortful listening (e.g., signal-to-noise ratio (SNR), spectral resolution, task performance). While previous research has provided broad insight into some of the more robust mechanisms that may underpin listening effort, it is limited in its clinical application. Therefore, the aim of this thesis is to explore how pupil dilation and changes in cortical oscillations in the alpha band are influenced by working memory capacity and task difficulty (by modulating both signal-to-noise levels and spectral resolution), whether they demonstrate similar behaviours

(suggesting that they may be markers of the same construct), and how different methodological aspects of data processing and statistical analysis might influence the outcomes and interpretation of a single dataset.

1.2 Background

Over the past few decades, listening effort has been conceptualised in multiple ways. Hicks and Tharpe (2002) defined listening effort as the attentional allocation required for speech understanding. This was largely based on early work by Downs (1982), which focused on hearing aid use and listening effort assessed using reaction times, and Feuerstein (1992), who examined how monaural and binaural hearing influenced perceived listening effort (assessed using a rating scale) and attentional effort (assessed using a dual-task paradigm). McGarrigle and colleagues (2014) proposed a dictionary-based definition: "the mental exertion required to attend to, and understand, an auditory message", and following expert consensus, the definition was further refined as: "mental effort as the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task that involves listening" which additionally captures how goal oriented behaviour may interact with effort allocation (Pichora-Fuller et al., 2016).[2]

Certainly, consensus suggests that listening to and recognising speech involves more than just hearing a sequence of sounds; it requires intent to extract meaning (Kiessling et al., 2003) and top-down processing needed when the auditory signal is not clear (Arlinger, Lunner, Lyxell, & Pichora-Fuller, 2009; Pichora-Fuller, Schneider, & Daneman, 1995). Successful speech recognition requires overlapping and interacting processes along the auditory pathway, including auditory processing and cognitive operations (Wingfield, Tun, & McCoy, 2005). In

[2] Based on the knowledge available at the time the experimental component of this thesis was developed (i.e., McGarrigle et al, (2014)) the stimuli and experiments were designed without factoring in how motivation can modulate listening effort.

addition, external factors may also influence listening effort. Examples of specific factors which could modulate listening effort include listener-internal and listener-external factors (from Lemke & Besser, 2016). Listener-internal factors encompass both auditory and cognitive processing, where auditory processing relates to individual ability (e.g., audiometric thresholds, spectral and temporal detection, and cognitive processing involving both linguistic ability and general cognitive mechanisms), whereas listener-external factors are the physical features that comprise a listening situation (e.g., reverberation, SNR, accented speech). Recently, a vectorized framework establishing how both listener-external and listener-internal factors influence listening effort in a resource limited system has been proposed (Strauss & Francis, 2017) which also encapsulates motivational factors discussed in Pichora-Fuller et al. 2016 (see forthcoming discussion in 1.3.2).

1.2.1 Listener-external factors

Across the majority of studies, listening effort is typically manipulated by varying the task demands related to listener-external factors, through speech reception thresholds (SRTs), SNRs, spectral or temporal degradation, different types of noise maskers, and linguistic complexity. By increasing task demands, it is assumed that the load on cognitive processes is increased, which may in turn modulate listening effort.

1.2.2 Listener-internal factors

When measuring listening effort, we are measuring how an individuals' hearing, cognitive and linguistic processes interact and respond to the demands of an auditory task. For speech recognition to be successful, three distinct, yet overlapping and interacting processes are required: 1) good peripheral hearing acuity; 2) intact central auditory processing; and 3) normal cognitive operations (Wingfield et al., 2005). Hearing impairment negatively affects speech audibility and the ability to discriminate between speech segments. Hearing devices, while effective at amplifying sound, often fall short of restoring the full spectral and temporal

properties of a speech signal required for successful speech discrimination/ recognition. While cognitive operations may compensate for peripheral and central auditory processing deficits (Wingfield et al., 2005), this process consumes cognitive capacity, limiting the availability of cognitive resources required for successful sentence processing (Lunner, Rudner, & Rönnberg, 2009; Mishra, 2014). Listening with a hearing impairment and/or in adverse conditions taxes cognitive resources, which modulates the amount of effort required to understand speech in difficult listening environments (Rabbitt, 1991; Pichora-Fuller, Schneider, & Daneman, 1995; Wingfield, Tun & McCoy, 2005; Rudner, Lunner, Behrens, Thorén, & Rönnberg, 2012).

Additionally, we are measuring an individual's motivation towards performing the task or achieving a specific outcome, presumably resulting in a non-linear relationshop between task demands and effort. The Framework for Understanding Effortful Listening (FUEL), discussed in detail in section 1.3.2, illustrates this three-way relationship (Figure 1; Pichora-Fuller et al., 2016).

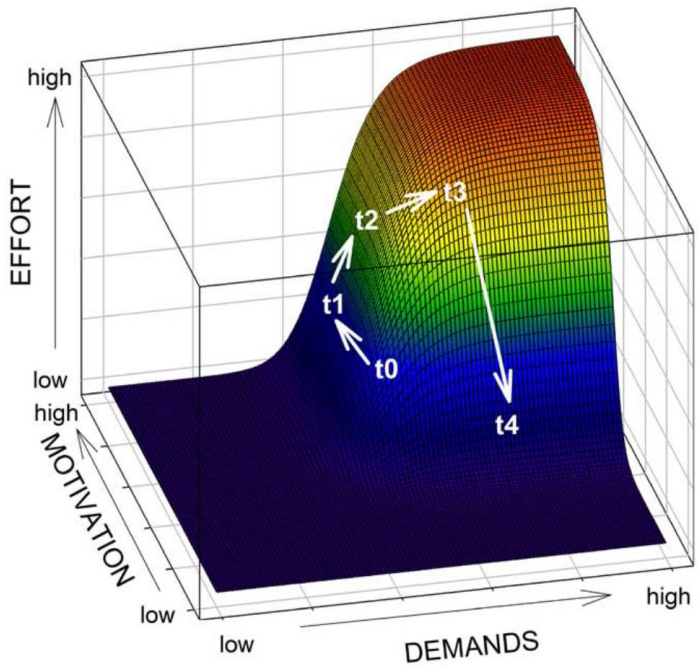

Figure 1. The 3-D representation of the FUEL illustrating the three-way relationship between listening effort, motivation and task demands from Pichora-Fuller et al. (2016). t0-t4 refer to time, discussed in more detail in Pichora-Fuller et al. (2016).

A 2007 review by Humes indicated that for hearing-impaired listeners, age and cognitive factors were the best determinants of speech recognition performance after controlling for audibility. Multiple cognitive operations have been implicated in speech understanding, including working memory capacity (Akeroyd, 2008), attention (Koelewijn, Shinn-Cunningham, Zekveld, & Kramer, 2014), and speed of processing (Schneider, Pichora-Fuller, & Daneman, 2010). Recently, Pichora-Fuller and colleagues (2016) suggested that evaluating these cognitive processes could potentially be used as indicators of listening effort.

1.3 Frameworks and models

Multiple frameworks and models exist to describe how auditory, cognitive and motivational factors might influence speech understanding and therefore, the magnitude of listening effort needed or invested during a task. The following provides a brief review of the main frameworks and models that will be discussed throughout the thesis. Note that while every effort is made to link the frameworks and models to the findings in this thesis, the Framework for Understanding Effortful Listening (FUEL) was put forward after the experiments of the current thesis were completed. As detailed below (section 1.3.2), a motivational component of listening effort is introduced in the FUEL. As such, this component was not manipulated in the experiments that form this thesis, and this limitation is discussion in Chapter 7.

1.3.1 The Ease of Language Understanding

The Ease of Language Understanding (ELU) model illustrates the function of working memory capacity in speech understanding in quiet and adverse conditions. Working memory may serve to support the integration of a received speech signal with both phonological and

semantic information stored in long term memory (Baddeley, 2003). Within the ELU framework, the process of integration occurs in a buffer called Rapid, Automatic, Multimodal Binding of PHOnology (RAMBPHO). If a mismatch occurs between a speech signal and a lexical representation in semantic long-term memory, explicit working memory processes are recruited to mediate the mismatch, and the process is delayed. Mismatches also result in increased working memory load as the item is kept longer in working memory whilst attempted mapping to lexical representations continues (Rudner, 2016). The extent to which working memory processes are involved in understanding speech is assumed to reflect the ease of listening (e.g., low working memory involvement when speech is highly intelligible) and listening effort (e.g., high working memory engagement when speech is highly unintelligible).

1.3.2 Framework for Understanding Effortful Listening (FUEL)

The FUEL modifies Kahneman's capacity model of attention as a framework to understand listening effort. Kahneman's original model (1973) describes attention as a limited capacity resource. The allocation of one's attention (such as attending to speech in multi-talker babble) is determined by various factors such as salience, compliance with task instructions, and willingness to complete a task, with arousal levels modulated by task and preparatory demands and "miscellaneous determinants" which include intensity of stimulation, drug interactions, and drive states (Kahneman, 1973). The two outputs in the model include "responses" and "miscellaneous manifestations of arousal" which include autonomic functions such as increased heart rate and pupil dilation.

The FUEL modifies and elaborates on Kahneman's model to make it more relevant to listening effort. Major changes include how (dis)pleasure may modulate motivation and performance, the renaming of the modules, and aligning examples with factors related to effortful listening such as spectral resolution and cognitive ability. The FUEL considers

motivation to be an important factor of listening effort. For instance, as task demands near impossible, an individual may not be motivated to perform given the effort-reward imbalance. On the one hand, some individuals may be highly motivated to perform well when task demands are high, and effort is mobilised to perform the task. The FUEL therefore considers how an interaction may exist between motivation and task demands and how this dynamic may modulate listening effort.

1.4 Methods used to assess listening effort

Listening effort has been assessed using a wide range of methods which fall into three broad categories: subjective, behavioural and physiological measures (Rudner et al., 2012). Subjective measures include anecdotal reports, questionnaires such as the Speech, Spatial and Qualities of Hearing Scale (Gatehouse & Noble, 2004), and rating scales such as visual analogue rating scales, the Borg Scale of Perceived Exertion (Borg, 1982), and the NASA-Task Load Index (Hart & Staveland, 1988). Behavioural measures consist of performance and/or response times during single or dual-task paradigms. The majority of physiological studies have focused on pupil dilation and neural oscillations, although various other measures including skin conductance, heart rate (measured as beats per minute and variability in rate), event related potentials (ERPs) and functional magnetic resonance imaging (fMRI) have been evaluated as potential tools to assess listening effort. While each of these measures have advantages and disadvantages which warrant consideration in developing a clinically viable tool to assess listening effort, this thesis limits its review to pupillometry and neural oscillations due to the breadth of research in these areas, and their clinical potential in regards to accessibility and affordability.

1.4.1 Subjective measures

Many studies have used self-report scales as a measure of perceived listening effort, however these are most often used to complement a corresponding objective measure of effort, such as

pupillometry or dual-task paradigms, rather than used as an independent measure. As there is no standardised or validated self-report scale of listening effort used across studies, there is limited ability to evaluate its clinical applicability (see Ohlenforst et al., 2017). For example, listening effort scales have disparate rating systems (Desjardins & Doherty, 2013; Hällgren, Larsby, Lyxell, & Arlinger, 2005; Hicks & Tharpe, 2002) and differ in their application, for example listening effort in some studies is rated after a single sentence but in others it is rated after a block of sentences. Despite this variability, as task difficulty increases (with changes in SNR, background maskers and/ or linguistic manipulations), ratings of effort increase, reflecting a commensurate change in perceived listening effort. Further, hearing-impaired listeners typically rate listening effort higher than those with normal hearing (Hua, Karlsson, Widén, Möller, & Lyxell, 2013; Luts et al., 2010), consistent with the change observed with other measures of listening effort (McCoy et al., 2005). See Klink, Schulte, and Meis (2012) for an in-depth review of subjective listening effort ratings.

Self-reported listening effort measures are generally easy to administer, cost effective, and do not require specialised equipment. However, their clinical applicability across the lifespan is limited. For example, young children may not understand what listening effort is and may not be able to effectively report it, older adults tend to under-report the extent of an impairment and over-report the extent of their ability (Uchida et al., 2003), whereas younger adults tend to over-report the extent of their difficulties (Seeman & Sims, 2015). Moreover, listening effort scales may not capture the right metric; some individuals may be inclined to rate their estimated performance on a task compared with the effort that was invested to perform the task (McGarrigle et al., 2014).

As well as self-report scales, questionnaires have been used to assess perceived listening effort. The Speech, Spatial and Qualities of Hearing Scale (SSQ; Gatehouse & Noble, 2004) comprises a sub-set of three questions pertaining to listening effort, with some studies

adapting the questions for experimental studies (Hornsby, 2013; Picou & Ricketts, 2014). Recently, Alhanbali, Dawes, Lloyd, and Munro (2017) developed the Effort Assessment Scale (EAS) which comprises 6 questions to assess perceived listening effort in everyday life (i.e., not during an experimental study). Questions range from "Do you have to put in a lot of effort to hear what is being said in conversation with others?" and "How easily can you ignore other sounds when trying to listen to something?". In validating the assessment, four groups of individuals aged between 55-80 years participated in the study which comprised hearing-aid users, cochlear implant users, those with single-sided deafness, and a normal hearing control group. Overall, the results showed that those groups with hearing impairment demonstrated significantly higher listening effort compared to the normal hearing control group. While such measures may have future relevance in the management of individuals with hearing loss, they are limited (like all retrospective questionnaires) to memory recall and to the relevance of questions asked, but are unlikely to be sensitive to evaluating the efficacy of different hearing devices, digital processing strategies, or therapies.

1.4.2 Behavioural measures

Behavioural measures of listening effort are typically evaluated using single / dual-task paradigms, with performance differences and / or reaction times as the outcome measure. Studies have typically found that when primary task difficulty is modulated through SNR adjustment, poorer SNRs lead to increased listening effort as indicated by performance decrements on the secondary task (Fraser, Gagné, Alepins, & Dubois, 2010; Gosselin & Gagné, 2011). There is a growing body of work investigating behavioural measures of listening effort, especially dual-task paradigms (see Gagné, Besser, & Lemke, 2017, for a comprehensive review of the field). Akin to self-reported measures, behavioural measures are cost-effective and do not require specialised equipment. They can, however, be challenging to administer and the complex task instructions may not be suitable for all age groups.

Individuals are also required to perform at their full potential throughout the duration of the task which can be difficult to monitor.

1.4.3 Physiological measures

Physiological measures of listening effort are direct measures of an individual's physical response to stimuli. These responses can be divided into central nervous system (CNS) and autonomic nervous system (ANS) activity (part of the peripheral nervous system). CNS activity has typically been assessed using non-invasive techniques such as encephalography (EEG), which detects voltage fluctuations from neural firing in the cortex, magnetoencephalography (MEG) which records the magnetic fields generated from neuronal firing, and fMRI which measures blood flow associated with neural activity. In the listening effort literature, M/EEG have been used to measure ERPs and changes in neural oscillations.

Neural oscillations arise from large groups of cortical neurons firing synchronously and can be classified in different frequency bands which have been shown to be modulated by a variety of cognitive processes (Başar, Başar-Eroglu, Karakaş, & Schürmann, 2001; Herrmann, Fründ, & Lenz, 2010; Klimesch, 1996, 2012; Klimesch et al., 1999; Ward, 2003). In comparison to other frequency bands, alpha waves (oscillations between 8 and 12 Hz; Klimesch, Doppelmayr, Pachinger, & Ripper, 1997) have received considerable attention in the listening effort literature. The origin of alpha oscillations remains relatively unclear, with early proposals theorising that the thalamus generated cortical oscillations by means of thalamo-cortical projections (Andersen, Andersson, & Lomo, 1968) leading to the view that the cortical oscillations were modulated by a 'pacemaker' (see Başar et al., 1997, for review). Subsequent evidence from alpha oscillatory studies demonstrated that this view may be too simplistic, as recordings at different cortical regions show marginally different central alpha frequencies (Lopes da Silva, van Lierop, Schrijer, & Storm van Leeuwen, 1973). Recently, it

has been suggested there are likely to be multiple autonomous alpha generators supporting disparate functional roles (Cohen, 2017).

There is evidence to suggest that changes in the alpha frequency band can be modulated by working memory load, and may reflect a functional inhibitory process (Jensen & Mazaheri, 2010; Klimesch, 2012). For example, in adapted auditory versions of the Sternberg paradigm, which requires participants to remember between two and six items and make a judgement after being prompted whether an item was presented in the list, alpha enhancement has been observed using syllables (Leiberg, Lutzenberger, & Kaiser, 2006), and single words (Karrasch, Laine, Rapinoja, & Krause, 2004; Pesonen, Björnberg, Hämäläinen, & Krause, 2006). Alpha power changes in an auditory spatial working memory task (Kaiser, Heidegger, Wibral, Altmann, & Lutzenberger, 2007), and in response to varying levels of acoustic degradation (Petersen, Wöstmann, Obleser, Stenfelt, & Lunner, 2015; Strauß, Wöstmann, & Obleser, 2014; Wöstmann, Herrmann, Wilsch, & Obleser, 2015) have also been demonstrated. Obleser, Wöstmann, Hellbernd, Wilsch, and Maess (2012) proposed that an acoustically degraded signal would exploit the same alpha oscillatory network as working memory load due to the greater allocation of working memory resources required to comprehend an acoustically degraded signal (Piquado, Isaacowitz, & Wingfield, 2010; Rabbitt, 1968; Wingfield et al., 2005). Using an auditory Sternberg paradigm, Obleser and colleagues (2012) parametrically varied working memory load (2, 4 and 6 to-be-remembered digits) and acoustic degradation (4, 8, and 16-channel vocoding) and showed a memory load and acoustic degradation-dependent alpha power enhancement over the central-parietal regions in both the encoding (stimulus presentation) and the delay (stimulus-free) periods. Interestingly, when both memory load and acoustic degradation were most challenging, alpha enhancement was super-additive, suggesting that acoustic degradation may draw on the same alpha network as working memory load. Similar findings have been replicated using different

sources of acoustic degradation during various comprehension tasks (Petersen et al., 2015; Wöstmann et al., 2015; see Strauß et al., 2014, for review).

Petersen et al. (2015) sought to investigate the effects of working memory load and acoustic degradation on the alpha oscillatory network when the auditory system was compromised by hearing impairment. Participant groups comprised older adults with normal hearing, mild hearing loss or moderate hearings loss. Controlling for audibility (using a fixed performance of 80% SRT) and age, clear spoken digits were embedded in speech-shaped background noise to vary acoustic degradation, and participants performed an auditory Sternberg task. Results showed that while participants with moderate hearing loss showed increased alpha power in comparison to those with lesser degrees of hearing loss in most conditions, under the highest working memory load combined with the greatest acoustic degradation, they displayed a decrease in alpha power. This was interpreted as reflecting neural break-down as a consequence of having to engage more working memory resources than their normal-hearing counterparts.

Much of the previous research on how acoustic degradation modulates alpha power has been conducted using closed-set digit stimuli, syllables, and words. Within the few studies examining sentence stimuli, changes in alpha power during a speech recognition task using channel-vocoded sentences (16- and 6-channel vocoding) and a vocoded 4-talker babble-noise varying from -7 dB to +7 dB SNR, McMahon et al. (2016) demonstrated that alpha power significantly declined with increasing SNRs only for the moderately easy 16-channel vocoded sentences. Conversely, the different SNR levels across the performance intensity function had relatively little impact on alpha power for 6-channel vocoded sentences.

The growing body of work on alpha power and listening effort lends initial support to its potential to be administered in a clinical environment, although much work is needed to

determine how different types of stimuli in clinical assessments may modulate alpha power. Importantly, it is also unknown how changes in alpha activity can provide information about the magnitude of listening effort invested by a single individual, as the published literature has focused on group-level analyses. Equipment such as the HEARLab® which is currently used to monitor the auditory brain response (ABR) and EEG in specialised audiology clinics, demonstrates that the measure has the potential to be easily integrated in a clinical environment, however MEG is relatively inaccessible due to it being highly specialised and expensive. M/EEG are also susceptible to an individual's movement, which could pose particular challenges for the young. Another potential challenge is that a large number of trials may need to be captured in order to separate the signal from the noise.

Compared with CNS assessments, the ANS is mostly unconsciously controlled and is responsible for regulating internal organ function such as blood pressure, heart and breathing rates, the production of bodily fluids (e.g., sweat), and digestion. In the listening effort literature, ANS activity has focused on skin conductance, heart rate measures, and pupillometry, which is a focus of this thesis.

Pupillometry is used to measure the diameter of the pupil. Stimulus-evoked pupil dilation is modulated by the release of norepinephrine (NE) from the locus coeruleus (LC; Wilhelm, Wilhelm, & Lüdtke, 1999) and has been implicated in prefrontal cortex activity (Laeng, Sirois, & Gredebäck, 2012). This LC-NE system has been hypothesised to mediate the entire attentional system (Corbetta, Patel, & Shulman, 2008) which has predominantly been assessed in visual-spatial based experiments using research methods from imaging and pharmacology in human subjects, and cellular recordings during animal experimentation.

Within the LC, cells show tonic or phasic modes of activity. Tonic activity slowly adapts to a stimulus-event, whereas phasic activity quickly adapts to a stimulus-event. In an unaroused

condition, tonic neuron firing is low which enables task and/or environmental disengagement (Aston-Jones & Bloom, 1981). Moderate activity is present during task focussed behaviour where reward is high (Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999). When engaged in an environment where there is doubt regarding an appropriate relationship between a stimulus-event and a response, tonic activity is high (Aston-Jones, Rajkowski, & Kubiak, 1997). The transition between tonic states is facilitated through dense cortical projections from the prefrontal cortex to the LC, sensitive to both task environment and reward (see Corbetta & Shulman, 2002, for review). During moderate tonic activity, phasic activity is at its highest, rapidly responding to a stimulus-event. This phasic response accelerates behavioural responses by boosting neural gain (Aston-Jones & Cohen, 2005), is associated with elevated task performance (Aston-Jones, Rajkowski, Kubiak, & Alexinsky, 1994; Bouret, Duvel, Onat, & Sara, 2003), supports selective attention (Aston-Jones, Rajkowski, & Cohen, 1999) and exhibits a reduction in baseline LC firing (conversely, tonic mode displays increased neuronal activity in the baseline region; Aston-Jones et al., 1994). Due to the pupil and LC-NE correlation, the pupil response can be exploited to uncover LC-NE activity (Koss, 1986) which may allow monitoring of attentional states (Laeng et al., 2012).

The pupil response has been used across multiple disciplines as a measure of exerted effort, and has been widely adopted to measure listening-related effort with currently upwards of 25 published studies. Typically, as task difficulty increases, the pupil diameter increases reflecting the increase of cognitive load. This metric has been suggested to index listening effort (Zekveld, Kramer, & Festen, 2011).

Pupil dilation and listening effort studies have mostly focussed on young normal hearing adults, with task difficulty modulated through SRT/ SNR adjustment and a further manipulation such as spatial differences (Koelewijn, de Kluiver, Shinn-Cunningham,

Zekveld, & Kramer, 2015; Koelewijn, Shinn-Cunningham, Zekveld, & Kramer, 2014; Zekveld, Rudner, Kramer, Lyzenga, & Rönnberg, 2014), different types of background maskers (Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; Zekveld & Kramer, 2014), channel-vocoding (McMahon et al., 2016; Zekveld, et al., 2014) and/or linguistic manipulations (Wendt, Dau, & Hjortkjær, 2016). The effect of hearing impairment on listening effort, measured through pupil dilation, has mostly been examined in older populations, again with SRT/ SNR modulation and a secondary adjustment such as the influence of different types of background noise (Koelewijn, Zekveld, Festen, & Kramer, 2014; Kuchinsky et al., 2014; Zekveld et al., 2011), and/ or linguistic manipulations (Kuchinsky et al., 2013; Kuchinsky et al., 2016), with few studies examining young children with hearing impairment, although see Steel, Papsin, and Gordon (2015).

Pupil dilation has been shown to increase with increasing task demands, even when performance accuracy is at, or near, 100% (Winn, Edwards, & Litovsky, 2015). Further to this, the pupil has been shown to dilate to a greater extent across different listening conditions when performance is comparable between the two (Koelewijn, Zekveld, Festen, & Kramer, 2012). Together, these results indicate that pupillometry has the potential to greatly increase the sensitivity of current clinical assessments, providing further insight into the listening challenges individuals face when performance accuracy is at ceiling, or comparable, across different listening conditions. This may greatly assist with intervention strategies, including device selection that provides optimal speech perception and ease of listening.

Pupillometric equipment is becoming more portable, user-friendly, and inexpensive, making it highly accessible for wide-clinical use. The large body of research into the pupil response to listening challenges is also encouraging for its potential as a clinical tool. Yet despite its research success, a practical clinical tool needs to be sensitive, and respond reliably, at the individual level. Further discussion regarding this is provided below.

1.4.4 Summary of measures

There are many advantages and disadvantages to applying a subjective, behavioural or physiological measure of listening effort in a clinical setting. Broadly speaking, subjective measures pose particular challenges in their clinical application due to participant biases and suitability across the lifespan. Similarly, task instructions on behavioural measures such as the dual-task paradigm may be too complex for some populations. However, both subjective and behavioural measures do not require specialised equipment and can be relatively easy to implement in the clinic. Conversely, physiological measures require specialised equipment, are not generally considered as part of routine audiological practice, set up can be time consuming, and the cost/ benefit may be discouraging. Advantages of physiological measures include their independence from rater biases, they do not rely on perceived invested effort which can be a difficult concept to understand, and are particularly beneficial for populations when behavioural responses are negligible or even non-existent (Wisniewski et al., 2015). Physiological measures of listening effort also provide a quantifiable metric which is largely used to determine candidacy for specialised intervention and insurance schemes (Dorman et al., 2012). Furthermore, with the right tool, it may be possible to combine a listening effort assessment into a pre-existing clinical assessment. For example, a physiological measure may be able to be assessed during a speech perception test which is already routinely conducted as part of a gold-standard assessment, therefore not requiring an additional assessment in an already time restricted schedule. There should, of course, also be minimal setup of the physiological tool in order to not interfere with time restrictions.

Even if a tool is suitable for the clinical environment, much work is needed to understand how different task parameters influence an individual's listening effort, how the measures behave across the lifespan, and their interactions with comorbidities. Further to this, different research groups vary in their data analysis and statistical modelling techniques making it

challenging to draw conclusions across the different studies. At present, best practices are lacking in this domain. Finally, there has been little to no discussion in the literature relating to intra- and inter-individual variability in relation to physiological measures. If a tool to assess listening effort is to be clinically implemented, understanding how the measures behave at the individual level is critically important.

1.5 Rationale for the thesis

The primary complaint of people with hearing impairment is that it is difficult to listen in noise. The increased effort required to comprehend speech in noise can lead to adverse health and psycho-social consequences such as stress and fatigue, and ultimately to maladaptive behaviours such as social withdrawal. Developing a clinical tool to assess listening effort may increase the sensitivity of current clinical assessments, permitting the client and clinician to make better informed decisions about device settings, device selection, and intervention/ rehabilitation strategies to ease listening in everyday life.

1.6 Research questions and justifications

In the current thesis, we explore some of the listener-internal and listener-external factors which modulate listening effort. The overarching aim of this thesis is to examine whether current measures identified as indexing listening effort are clinically viable.

The specific aims of this thesis were to:

1) Examine whether working memory capacity interacts with subjective (Chapter 3) and physiological measures (Chapter 5) of listening effort,

2) Investigate how subjective (Chapter 3) and physiological measures (Chapter 4) of listening effort are modulated by spectral resolution (channel-vocoding) and background noise,

3) Determine whether pupil dilation and alpha oscillations - two physiological measures claimed to index listening effort - are correlated (Chapter 4), and

4) Investigate how differences in data processing and statistical approach can affect the results and interpretation of results in a single dataset (Chapter 6).

Chapter 2 General methodology

2.1 Participants

Participants were recruited through Macquarie University and were selected based upon the following criteria: aged between 18-35-years-old, monolingual English speaking, normal hearing, right handed, and no known neurological disorders. Normal hearing was determined by presence of distortion-product otoacoustic emissions between 1 – 4 kHz. Handedness was assessed by The Assessment and Analysis of Handedness: the Edinburgh Inventory (Oldfield, 1971). The Macquarie University Human Research Committee approved the studies.

All participants recruited for Chapter 3 Study (a) were asked to participate in a follow-up study (Chapter 4). The data collected from both Chapter 3 and Chapter 4 were used for subsequent analyses in Chapter 5 and 6. Table 1 shows the mean and SD of the participant's ages. The varying sample sizes across studies reflect the data available depending on the measures being analysed. The participants recruited for Chapter 3 Study (b) were not included in any further chapters/ analyses.

Table 1. Number and age of participants included in the final analyses of each chapter.

Chapter number	n	Age (mean)	Age (SD)
3 Study (a)	24 (13F)	27.3	4.5
3 Study (b)	20 (13F)	22.7	2.8
4	19 (12F)	27.0	4.3
5	19 (12F)	27.0	4.3
6	23 (13F)	27.5	4.0

2.2 Study design

All papers used the same stimulus materials, and background noise.

To investigate changes in perceived and physiological measures of listening effort, spectral resolution and background noise were parametrically varied. Spectral resolution was manipulated through channel-vocoding and comprised two conditions: moderately challenging (6-channel vocoding), and less challenging (16-channel vocoding). Channel vocoding sentences provides, in a controlled manner, a method to reduce the spectral resolution of the signal whilst maintaining temporal information. In doing so, speech becomes more challenging to understand with decreasing channels (Friesen, Shannon, Baskent, & Wang, 2001). We therefore anticipate the amount of listening effort required to comprehend a sentence in the 6-channel condition will be greater than in the 16-channel condition.

In order to maintain performance levels across participants, background noise across all studies was adjusted for each participant to obtain 50% (moderately challenging) and 80% speech reception thresholds (SRTs; less challenging). Note for Chapter 3, Study (b), a fixed 50% SRT was first determined, and +/-3 dB was added to each participants' 50% SRT. Fixed performance levels using SRTs were chosen instead of fixed signal-to-noise ratios (SNRs) as cognitive factors may modulate performance when using fixed SNRs (Souza & Arehart, 2015). Moreover, we speculated that results from our previous study using fixed SNRs (McMahon et al., 2016) may have been due to differences in performance, and therefore wanted to investigate whether controlling performance results in the same outcome. We anticipated that listening to a sentence at a 50% SRT will be more effortful than listening to sentences when performance is higher (e.g., 80% SRT, and the +3 dB condition in Chapter 3, Study (b).

2.3 Stimulus materials

2.3.1 Sentences

BKB sentences adapted for Australian-English (Bench & Doyle, 1979) were recorded by a native Australian-English female speaker. The sentences were vocoded using custom MATLAB scripts. The frequency range was set to 50-6000 Hz and divided into either 6 or 16 logarithmically-spaced channels. The amplitude envelope was then extracted from each channel by full-wave rectifying the signal and applying a low pass filter with a 32 Hz cut-off frequency. The extracted envelope was used to modulate the noise with the same frequency band. Each band of noise was then recombined to produce the channel-vocoded sentences. The root mean square (RMS) levels of the sentences were equalised in MATLAB after vocoding.

2.3.2 Background noise

The background noise was 4-talker babble, which is the speech-noise used within an Australian clinical context. The same channel-vocoding and RMS adjustment used for the sentence materials was applied to the background noise.

2.4 Cognitive measures

Working memory capacity

Participants were presented with consecutive lists of 3, 4 and 5 sentences, respectively. Each sentence comprised 3-5 words and were either normal sentences (e.g., The girl hugged the father), or absurd (e.g., The bear wrote poetry) and were grammatical or agrammatic. Sentences appeared on the laptop monitor one word at a time and when the sentence was completed, there was a short pause where the participant was required to answer whether the sentence made sense or not by pressing the Y or N key. After a block of sentences were presented (e.g., 1 x 3 sentences), the participants were asked to repeat aloud either the first words (using the examples above, the correct answers would be ‘The girl’ and ‘The bear’, or

the final words ‘the father’ and ‘poetry’). These answers were then recorded by the examiner. The aggregate score of the entire task was then used to calculate the percentage out of 24.

Chapter 3 Perceived listening effort ratings and the influence of working memory capacity

Kelly Miles[1,2,3], Catherine McMahon[1,2], Isabelle Boisvert[1,2], and Björn Lyxell[3]

[1] Centre for Implementation of Hearing Research, Macquarie University, Sydney, Australia

[2] The HEARing Cooperative Research Centre, Melbourne, Australia

[3] Linnaeus Centre for HEaring And Deafness (HEAD), Swedish Institute for Disability Research, Linköping University, Sweden

3.1 Abstract

This research aims to investigate whether working memory capacity influences perceived listening effort ratings. Across two studies, 43 normal-hearing adults (18-35 years) participated in a speech recognition task, in which speech stimuli were degraded by manipulating both the signal-to-noise ratio and the level of channel-vocoding. Performance on the task was measured and using a modified Borg scale, participants were asked to rate how effortful they found the task in each condition. A reading span task was used as a measure of working memory capacity and participants were stratified into a high or low working memory capacity group using a median split across the group. Data from both studies showed that working memory capacity influences perceived listening effort ratings. Specifically, in contrast to the group with higher working memory capacity, those with lower working memory capacity did not differentiate perceived effort ratings across speech reception thresholds when spectral resolution was greatest. These results suggest that working memory capacity influences perceived listening effort ratings. The disparity in perceived listening effort ratings for the different working memory capacity groups may be related to differences in signal adaptation, and/ or inhibitory factors. Clinical applicability and suggestions for future research to better understand these factors are proposed.

3.2 Introduction

Listening in sub-optimal conditions, with multiple talkers or in reverberant environments, can be demanding for young adults with normal hearing. For people with hearing loss however, listening in adverse conditions is particularly challenging and is a commonly reported complaint (Arlinger, Lunner, Lyxell, & Pichora-Fuller, 2009; Hawkins & Yacullo, 1984; Wouters & Berghe, 2001). Cognitive resources, such as working memory, may be taxed when listening to degraded speech, and this may increase the amount of effort required to understand speech in difficult listening environments (Rabbitt, 1991; Wingfield, Tun & McCoy, 2005). Older adults with hearing loss may be further disadvantaged in understanding speech in noisy environments by the combined effects of sensory and cognitive declines (see Pichora-Fuller & Singh, 2006 and Wayne & Johnsrude, 2015). However, as some cognitive abilities such as vocabulary improve with increasing age (Park & Reuter-Lorenz, 2009), there may exist a performance trade-off between factors such as slowed cognitive processing speed and increased vocabulary and general knowledge that comes with aging. Individuals do however vary in their cognitive capacity and in their social and physical environments. While hearing loss itself may contribute to increased listening effort required to understand speech in adverse environments, it is important to understand the effects that listening effort has on the individual, and their ability to communicate and fully participate in society.

The World Health Organisation's (WHO) International Classification of Functioning, Disability and Health (ICF) provides a framework to address the bio-psychosocial factors that contribute to disability, including the interaction of health conditions (in this case hearing loss) with personal and environmental factors that may influence participation (World Health Organization, 2001). Sustained effort could lead to multiple negative health consequences, including increased stress (Hua et al., 2014), cardiovascular strain (Peters et al., 1998) or fatigue (Mehta & Agnew, 2012). For people with hearing loss, anecdotal (Kramer, Kapteyn,

& Houtgast, 2006; Pichora-Fuller, 2003) and interview reports (Hua, Anderzén-Carlsson, Widén, Möller, & Lyxell, 2015) indicate that fatigue may be linked to listening in noisy environments, potentially due to the effort expended when listening with compromised audition (Edwards, 2007). This may, at least in part, account for people with hearing loss taking more sick-leave (Kramer et al., 2006) and requiring a longer time for recovery after work (Nachtegaal et al., 2009). Further, increased effort during listening could reduce a person's ability to engage in cognitive dual-tasks which might lead to adverse events, such as increased risk of falling (Lin & Ferrucci, 2012) or work-related accidents (Picard et al., 2008).

When the auditory input that reaches the brain is compromised by degradation of the acoustic signal or a hearing impairment, the role of cognitive processes in understanding speech becomes increasingly important (Arlinger et al., 2009). In a review of 20 studies examining the relationship between speech recognition in noise and various measures of cognition, Akeroyd (2008) identified that working memory capacity measured using a visual reading span paradigm was the best predictor of speech recognition performance. For example, Lunner (2003) showed a strong significant correlation between a similar reading span task and speech reception thresholds (SRTs) where individuals with higher working memory had better performance outcomes – a finding which has since been replicated (Foo, Rudner, Rönnberg, & Lunner, 2007). As a decrease in cognitive resources may contribute to an increase in the amount of listening effort required to comprehend an incoming speech signal (Pichora-Fuller, Schneider, & Daneman, 1995; Rudner et al., 2011), the interplay between listening effort, working memory capacity and speech performance requires further investigation.

Working memory involves both the simultaneous storage and processing of relevant information over a short period of time (Baddeley, 2012; Daneman & Carpenter, 1980). This

ability is necessary for carrying out everyday tasks, and is limited in both time and capacity (Baddeley, 2012). Working memory capacity is a cognitive component essential in most communicative tasks (Lyxell et al., 2008) and may be taxed by background noise, even when the noise is not attended to (Gisselgård, Petersson, & Ingvar, 2004; Miles, Jones, & Madden, 1991; Salamé & Baddeley, 1989). Working memory processes may be engaged to distinguish meaningful speech material from task-irrelevant background noise (Bregman, 1994), and by top-down processing used to facilitate verbal inferences when segments of a message are lost or ambiguous due to signal degradation (i.e., 'filling in the gaps'; Boothroyd & Nittrouer, 1988; Flynn & Dowell, 1999; Grant & Seitz, 2000; Pichora-fuller, 2006; Wingfield, 1996). It is likely that increased effort associated with listening in noise is in part attributable to these types of cognitive processes.

The concept that listening effort ratings may be derived from the relationship between cognitive processes and speech recognition performance was proposed by Rudner et al. (2011) and Rönnberg et al. (2013). The Ease of Language Understanding (ELU) model provides a framework for understanding how challenging listening situations can give rise to effortful listening (discussed in detail in 1.3.1). In brief, working memory capacity may support signal processing, with mismatches of phonological and semantic mappings drawing on explicit working memory processes to temporarily store items until successful mapping occurs. In demonstrating this, Rudner and colleagues (2012) controlled for working memory capacity's influence on listening effort ratings by using it as a covariate in an analysis of variance to examine this relationship. The results showed that listening effort was rated higher in the more challenging SNRs, however controlling for working memory (measured using a letter monitoring or reading span task) did not explain the variance in listening effort ratings. On the other hand, working memory did influence listening effort ratings in the different noise types where modulated noise was rated more effortful than steady-state noise.

The parameter estimates showed that individuals with higher working memory rated listening effort lower than those with lower working memory. Therefore, it is possible that individuals with higher working memory capacity may benefit from more efficient processing (Rudner et al., 2012).

Despite improved understanding of the link between cognition and speech understanding, audiological assessments used to evaluate speech recognition do not typically consider the role of cognitive processes in the management of hearing loss, nor its contribution to inter-individual differences in speech perception. As cognitive processes may contribute to individual performance and assist in explaining differences in perceived listening effort, it is important to quantify its role during speech recognition. This paper evaluates the influence of working memory capacity on listening effort ratings in varying levels of noise and spectral resolution, in normal-hearing young adults. Two studies were conducted in parallel (different participants in each study) to determine how changes in perceived listening effort were influenced by fixed performance (Study a: 50, 80% SRT), and relative performance (Study b: 50% SRT +/-3 dB). To investigate changes in perceived listening effort, spectral resolution and background noise were parametrically varied. Spectral resolution was manipulated through channel-vocoding and comprised two conditions: moderately challenging (6-channel vocoding), and less challenging (16-channel vocoding). Channel vocoding sentences provides, in a controlled manner, a method to reduce the spectral resolution of the signal whilst maintaining temporal information. In doing so, speech becomes more challenging to understand with decreasing channels (Friesen, Shannon, Baskent, & Wang, 2001). We therefore anticipate the amount of listening effort required to comprehend a sentence in the 6-channel condition will be greater than in the 16-channel condition.

In order to maintain performance levels across participants, background noise in Study (a) was adjusted for each participant to obtain 50% (moderately challenging) and 80% speech

reception thresholds (SRTs; less challenging). In Study (b), a fixed 50% SRT was first determined, and +/-3 dB was added to each participants' 50% SRT. Fixed performance levels using SRTs were chosen instead of fixed signal-to-noise ratios (SNRs) as cognitive factors may modulate performance when using fixed SNRs (Souza & Arehart, 2015). We anticipate that listening to a sentence at a 50% SRT will be more effortful than listening to sentences when performance is higher (e.g., 80% SRT, and the +3 dB condition in Study (b)).

Modulating both SRT and channel-vocoding permits comparison of how the two manipulations influence listening effort. SRT allows comparison of how performance relates to listening effort, and channel-vocoding further allows comparison of how signal resolution influences listening effort. Manipulating both factors is important in understanding the multidimensionality that comprises listening effort.

The motivation for conducting the parallel studies was to compare similarities and differences between fixed performance listening effort ratings (time consuming to clinically administer) and relative performance listening effort ratings (quicker to clinically administer). If the results are similar, this would suggest that the quicker and simpler relative performance procedure may be a better clinical assessment tool than the more time consuming fixed performance procedure. Gaining a better understanding of the interplay between listening effort, speech recognition, and cognitive processes may have implications for people with hearing impairment, such as device selection, setting selection, choice of cognitive intervention and rehabilitation strategies.

3.3 Method (Study a)

Participants

Twenty-four monolingual Australian-English speakers (13 women, 11 men) with a mean age of 27.3 years (SD = 4.5) participated in the study. All had distortion-product otoacoustic

emissions within normal limits between 1 – 4 kHz, consistent with typical hearing or a mild sensorineural loss only.

Speech materials and masker

BKB sentences adapted for Australian-English (Bench & Doyle, 1979) were recorded by a native Australian-English female speaker. The sentences and background noise (4-talker babble) were vocoded using custom MATLAB scripts where the frequency range 50-6000 Hz was divided into 6 or 16 logarithmically spaced channels. The amplitude envelope was then extracted from each channel by full-wave rectifying the signal and applying a low pass filter with a 32 Hz cut-off frequency. The extracted envelope was used to modulate the noise within the same frequency band. Each band of noise was then recombined to produce the channel-vocoded sentences and background noise. Channel-vocoding the background noise to match the speech signal was necessary in order to match the acoustic features of the signal to the background noise to minimise any cues provided by greater or lesser contrasting differences between them. The root mean square (RMS) levels of the sentences and background noise were equalised in MATLAB after vocoding.

An automated adaptive speech-in-noise software developed by the National Acoustic Laboratories was used to obtain SRTs (see Keidser, Dillon, Mejia & Nguyen, 2013, for a comprehensive review of the algorithm). The BKB-A sentences were presented at 65 dB SPL, and the background noise was adaptively adjusted to obtain each of the SRTs. The adaptive procedure consisted of three phases. Phase 1: 5 dB steps until 4 sentences were completed, including one reversal, Phase 2: 2 dB steps until a minimum of 4 sentences were completed, and the phase's standard error (SE) was 1 dB or below, and Phase 3: 1 dB steps until 16 sentences (from the end of phase 2) were completed, with a SE of 0.80 or below, or the maximum number of 32 sentences was reached (note that the minimum was 16

sentences). When the SE reached 0.80 or below, the test terminated and recorded the SNR. This procedure was conducted across both 50% and 80% SRTs.

Speech reception thresholds

A sound-attenuated room was used during the testing sessions, and the equipment was calibrated prior to each participant's arrival. The speaker was positioned at one meter distance and zero degrees azimuth from the participant at ear level. Two SRTs (50%, 80%) for two vocoding conditions (16-channel, 6-channel) were collected, resulting in four conditions. Table 1 shows the mean and standard deviations of the signal-to-noise ratios for each SRT.

Table 1. Mean signal-to-noise ratios and standard deviations for speech reception thresholds (SRTs: 50, 80%) and channel-vocoding (16, 6-channel).

SRT %	*50*		*80*	
Vocoding	*6*	*16*	*6*	*16*
Mean (dB)	1.0	-1.8	4.0	0.5
SD (dB)	1.9	1.5	2.1	1.6

Perceived effort ratings

Perceived effort ratings were obtained for all conditions. At the individual's SRT, one list of 16 sentences was presented to obtain perceived listening effort ratings using the Borg Scale of Perceived Exertion (Borg, 1982). The scale is ranked one to ten, with corresponding written labels (for example, '1' corresponded with the label 'Nothing at all', and '10' corresponded with the label 'Very very hard'). This scale differs from visual analogue scales due to the inclusion of written labels. The order in which the conditions were presented was randomised across participants.

Working memory capacity

The reading span test was used as a measure of working memory capacity (Daneman & Carpenter, 1980). Short sentences were presented visually on a computer screen, in three meaningful segments, each separated by a 50 ms gap (e.g., “the dad”, “hugged”, “the daughter”), in blocks of three, four, and five sentences. Each segment appeared for 800 ms. After each sentence, the participant had 1.75 seconds to determine whether the sentence made sense or not by pressing ‘Y’ or ‘N’ on a keyboard. After each block of sentences, participants were asked to repeat either the first or the last words of each sentence. The order of presentation was randomised. Scoring was based on the total number of words correctly recalled (Rönnberg, Arlinger, Lyxell, & Kinnefors, 1989).

3.4 Results (Study a)

A median split was performed on the basis of working memory capacity (visual reading span score) performance. The median was 67%, the mean of the lower and higher scores were 58% and 76%, respectively. Lower/ higher working memory capacity was entered into a repeated measures ANOVA as a between-subjects factor to determine how working memory capacity influenced the model. There was a significant main effect of SRT on effort ratings, $F(1,22) = 27.171$, $p < 0.001$, $\eta_p^2 = 0.553$, no main effect of vocoding, $F(1,22) = 0.004$, $p = 0.947$, $\eta_p^2 < 0.001$, or SRT x vocoding, $F(1,22) = 4.206$, $p = 0.052$, $\eta_p^2 = 0.161$, although this was trending towards significance (Figure 1). For low/ high working memory capacity, there was no interaction with SRT, $F(1,22) = 0.006$, $p = 0.938$, $\eta_p^2 < 0.001$, or vocoding $F(1,22) = 0.279$, $p = 0.602$, $\eta_p^2 = 0.013$. There was a significant interaction between SRT, vocoding and working memory capacity, $F(1,22) = 4.748$, $p = 0.040$, $\eta_p^2 = 0.178$ (Figure 2). Bonferroni corrected post-hoc analyses showed that on average, participants with higher working memory capacity rated effort higher in the 50% SRT condition compared to the 80% SRT condition, for both vocoding conditions. On average, participants with lower working memory capacity rated effort higher in the 50% SRT condition compared to the 80% SRT

condition for 16-channel vocoding only. Perceived effort was rated the same level in the most challenging vocoding condition (6-channel).

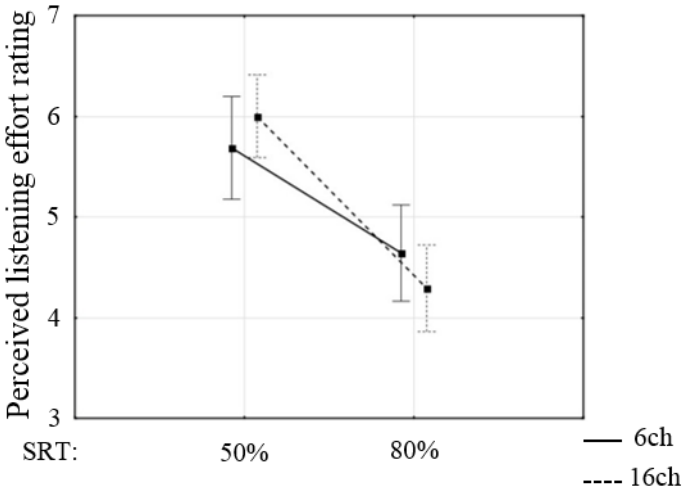

Figure 1. Interaction between perceived listening effort ratings, speech reception thresholds. (SRTs: 50%, 80%) and channel-vocoding (16, 6-channel). Error bars represent ±1SD.

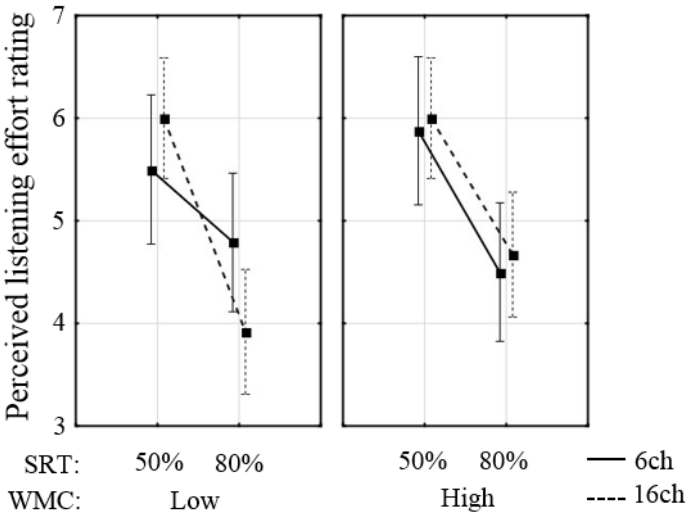

Figure 2. Interaction between perceived listening effort ratings, speech reception thresholds (SRTs: 50%, 80%) and channel-vocoding (16, 6-channel), by working memory capacity

(WMC: low, high). Error bars represent ±1SD.

Simple linear regression was calculated to predict perceived listening effort based on SNR, across all conditions. There were no significant results, suggesting that on average participants' perceived effort ratings were not correlated with SNRs (Table 2).

Table 2. P-values of simple linear regression where signal-to-noise ratios were used to predict perceived listening effort.

SRT %	*50*		*80*	
Vocoding	*6*	*16*	*6*	*16*
P-value	0.5	0.8	0.4	0.6

3.5 Method (Study b)

Participants

Twenty monolingual Australian-English speakers (13 women, 7 men) with a mean age of 22.7 years (SD = 2.8) participated in the study. All had distortion-product otoacoustic emissions within normal limits between 1 – 4 kHz, consistent with typical hearing or a mild sensorineural loss only.

Speech materials and noise

The same speech and noise materials were used as Study (a).

Speech reception thresholds

SRTs were collected using the same adaptive method as Study (a). A 50% SRT was first obtained for 16-channel and 6-channel vocoding conditions, then -3 dB and +3 dB was added to each participants' 50% SRT resulting in 6 conditions. Table 3 shows the mean and standard deviations of the SNRs for Study (b).

Table 3. Mean signal-to-noise ratios and standard deviations for speech reception thresholds (SRTs: +/- 3 dB from the participants' individually adapted 50% SRT) and channel-vocoding (16, 6-channel).

SRT %	*-3dB*		*50*		*+3dB*	
Vocoding	*6*	*16*	*6*	*16*	*6*	*16*
Mean (dB)	0.1	-4.0	3.1	-1.0	6.1	2.0
SD (dB)	2.4	1.7	2.4	1.7	2.4	1.7

Perceived effort ratings

Perceived effort ratings were obtained for all conditions, as in Study (a). At the individual's SRT, one list of 16 sentences was presented to obtain the effort ratings for the 50% SRT condition, and 32 sentences were presented in the 50% SRT -3 dB and +3 dB condition. The order in which the conditions were presented was randomised across participants.

Working memory capacity

The same working memory task was used as Study (a).

3.6 Results (Study b)

A median split was performed on the basis of working memory capacity (reading span score). The median was 71%, the mean of the lower and higher scores were 60% and 71%, respectively. Lower/ higher working memory capacity was entered into a repeated measures ANOVA as a between-subjects factor. There was a significant main effect of SRT on effort ratings, $F(2,36) = 26.890$, $p < 0.001$, $\eta_p^2 = 0.599$, no main effect of vocoding, $F(2,36) = 0.020$, $p = 0.887$, $\eta_p^2 = 0.001$, or SRT x vocoding, $F(2,36) = 2.763$, $p = 0.076$, $\eta_p^2 = 0.133$, although this was approaching significance (Figure 3). For low/ high working memory capacity, there was no interaction with SRT, $F(2,36) = 0.393$, $p = 0.677$, $\eta_p^2 = 0.021$, or

vocoding $F(2,36) = 1.014$, $p = 0.327$, $\eta_p^2 = 0.053$ and no significant interaction between SRT, vocoding and working memory capacity, $F(2,36) = 1.928$, $p = 0.160$, $\eta_p^2 = 0.096$.

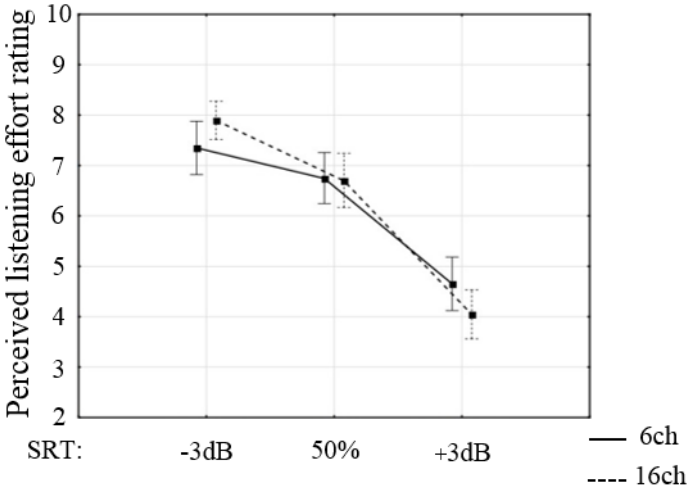

Figure 3. Interaction between perceived listening effort ratings, speech reception thresholds (SRTs: +/- 3 dB from the participants' individually adapted 50% SRT) and channel-vocoding (16, 6-channel). Error bars represent ±1SD.

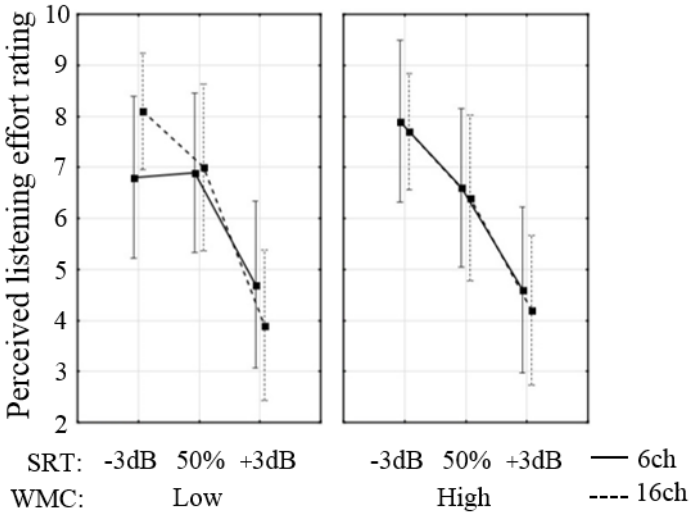

Figure 4. Interaction between perceived listening effort ratings, speech reception thresholds (SRTs: +/- 3 dB from the participants' individually adapted 50% SRT) and spectral resolution (vocoding: 16, 6-channel), by working memory capacity (WMC: low, high). Error bars represent ±1SD.

Two additional conditions (+/- 3 dB) for 16- and 6-channel vocoding conditions were calculated from the adaptively obtained 50% SRT condition. The actual performance levels attained (average percentage correct) are shown in Table 4. Descriptive statistics show that the difference between the fixed 50% SRT condition and the relative +3 dB conditions (for both vocoding conditions) had notable performance increases. Conversely, performance appears to be comparable between the -3 dB condition and the fixed 50% SRT condition.

Table 4. Mean and standard deviations for the actual performance levels obtained during the speech recognition task in Study (b), at each speech reception threshold (SRTs: +/- 3 dB from the participants' individually adapted 50% SRT) and channel-vocoding level (16, 6-channel), by working memory capacity (WMC: low, high).

WMC	*Low*						*High*					
SRT %	*-3dB*		*50*		*+3dB*		*-3dB*		*50*		*+3dB*	
Vocoding	*6*	*16*	*6*	*16*	*6*	*16*	*6*	*16*	*6*	*16*	*6*	*16*
Mean (%)	49	46	49	53	78	85	44	40	45	52	74	79
SD (%)	17	17	13	17	9	13	15	21	13	18	13	7

Simple linear regression was calculated to predict perceived listening effort based on SNR, across all conditions. No results were significant except for the 50% SRT 16-channel vocoding condition, suggesting that overall participants' perceived effort ratings were not correlated with SNRs (Table 5).

Table 5. P-values of simple linear regression where signal-to-noise ratios were used to predict perceived listening effort.

SRT %	*-3dB*		*50*		*+3dB*	
Vocoding	*6*	*16*	*6*	*16*	*6*	*16*
P-value	0.2	0.1	0.1	0	1	0.2

3.7 Discussion

Across two studies, a sentence recognition task which manipulated SNRs and channel-vocoding was conducted to examine how working memory capacity influences self-reported listening effort ratings in young, normal-hearing adults. The main finding was that on average, participants with lower working memory capacity rated perceived effort differently to the group with higher working memory capacity (Study a). This difference appeared to be driven by the interaction between SRTs and spectral resolution (channel-vocoding), and was only apparent in the more challenging vocoded condition (6-channel). That is, the lower working memory capacity group, on average, did not rate effort differently across the easier and harder SRTs when the vocoded signal was more challenging. In contrast, the higher working memory capacity group, on average, systematically rated effort higher in the more challenging SRT for both vocoding conditions, in line with pupillometry studies showing that on average, participants with greater working memory capacity exert greater effort than those with lower working memory capacity (Zekveld, Kramer & Festen, 2011). Although not significant, a similar trend was found in Study (b) (Figure 4).

One explanation is that the higher working memory capacity group were less distracted by the background noise than the lower working memory capacity group. Conway et al. (2001) reports that individuals with higher working memory capacity may have a greater ability to supress interference than their lower working memory capacity counterparts. Studies by Rosen and Engle (1998) and Conway and Engle (1994) found that participants with a higher working memory capacity were superior at suppressing interfering materials than those individuals with lower working memory capacity. In particular, in a study investigating the cocktail party effect, Conway, Cowan, and Bunting (2001) demonstrated that individuals with low working memory capacity were more distracted by hearing their own name presented

within cocktail party noise than those with higher working memory capacity. If this was the case in the current study, it may be that the group with higher working memory capacity were less distracted by the background noise as they had more resources (or 'cognitive spare capacity'; see Rudner et al., 2011) to suppress the distracting information. This may have the added benefit of allowing the participants to rate the effort associated with listening to the signal (sentence material), as opposed to rating the distracting background noise (or the combination of the two).

An alternate explanation is that differences exist in the ability of those with higher and lower working memory capacity to adapt to a degraded signal. Evidence that individuals with higher working memory capacity are better able to adapt to different signal processing algorithms than those with lower working memory capacity has been shown previously (Lunner, 2003; Lunner, Rudner, & Rönnberg, 2009; Rudner, Foo, Rönnberg, & Lunner, 2009). Lunner et al. (2009) showed that adults with hearing loss who have higher working memory capacity are better able to use faster signal processing algorithms in hearing aids than those with lower working memory capacity. Within the current study, young adults needed to correctly repeat blocks of vocoded sentences that were increased in difficulty either with increased 4-talker babble noise or a reduction in spectral content. Adaptation to a degraded signal over repeated trials has been shown by Davis and colleagues (2005), however it is noted that they did not include a measure of working memory capacity in their study. They found that young adults increased the proportion of correctly repeated key words in a sentence from less than 10% to approximately 50% from the first sentence in a block of 30 vocoded sentences to between 20-30 sentence presentations. As feedback was not provided to the participants, it is assumed that the increase in correctly reported words over the number of sentence presentations suggests that participants were able to adapt to the vocoded signal (or learnt to 'decode' the signal). To extend this, it is certainly possible that

differences could exist for adults with higher working memory, where they are able to learn or decode degraded sentences more rapidly, although this hypothesis has not yet been tested.

Despite the fact that perceived effort ratings were comparable across Study (a) and (b), variability in the actual performance levels obtained in Study (b) existed (Table 4). It would be expected that there would be greater variability in performance in the -3 dB and +3 dB conditions as the SNR was not adapted to reach an SRT (i.e., -3 dB and +3 dB were simply added to each individuals' 50% SRT). However, the results showed that considerable variability was present across all conditions, for both the lower and higher working memory capacity groups. Keidser and colleagues (2013) tested the speech-in-noise algorithm in a free-field (the current study also used the same software in a free-field), observing that variability across studies may arise due to free-field presentation where an individual's head movement may result in shadowing effects. Moreover, Dillon (1982) outlined that the variability often found in inter- and intra-individual results of speech discrimination tasks may arise from multiple variables such as list differences, statistical fluctuations and subject differences. In particular, time of day has been shown to affect performance and effort ratings in young adults. Recent work by Veneman and colleagues (2013) showed young individuals performed significantly better when they were tested at their 'peak-time' (evening) compared to in the morning, and their ratings of mental effort (captured using the NASA-TLX) were significantly higher when they were assessed during their off-peak time (morning). It is plausible that if all testing sessions were conducted during the participants' peak-time, performance variability would be reduced.

It was also expected that the lower working memory capacity group would rate effort higher than the higher working memory capacity group, however the results of the current study do not support this. Rudner et al. (2012) demonstrated across two studies that older individuals (M = 63.5; 70 years) with hearing-impairment and lower working memory capacity rated

effort higher than those with higher working memory capacity. One possible explanation for the inconsistency between effort ratings in Rudner and colleagues' study and the current study is likely due to the notable differences in age and hearing status of the participants (Wayne & Johnsrude, 2015). It is conceivable that in a young, normal-hearing population, effort ratings during a speech recognition task, and working memory capacity, would be more homogenous than in an elderly population where hearing-impairment and cognitive decline co-occur (Grady, 2012; Salthouse, 2004). It is also likely that the differences result from the sentence recognition task materials themselves. Studies have shown that elderly participants benefit more from context than their younger counterparts (Pichora-Fuller et al., 1995) and the study by Rudner and colleagues used low redundancy sentence stimuli. These materials essentially impede the ability of the participants to use contextual inference, whereas the current study used predictable sentences, allowing the participants to use context to 'fill in the gaps'. This may also explain why the older participants in Rudner et al.'s study rated perceived effort higher (1.46 times higher for the comparable 50% SRT conditions) than the younger participants in the current study.

Despite past research into listening effort, many challenges remain for a behavioural measure of listening effort to be implemented in a clinical setting. For example, further investigation is needed to understand the effect of personality or internal factors such as motivation on self-reported listening effort (Picou & Ricketts, 2014), and performance outcomes (Humphreys & Revelle, 1984). And while subjective effort ratings may in part reflect an individual's perception of listening effort, it may also be related to the actual or perceived performance level (Feuerstein, 1992). For example, recalling a sentence embedded in noise may require effort, but if the individual believes they have recalled the sentence successfully they may rate perceived effort lower, compared to a less effortful sentence (i.e., due to a favourable SNR) that they were either unable to recall in whole or in part, or unsure of having heard the

sentence correctly. Studies examining listening effort during a dual-task paradigm have shown that the subjective and objective measures of listening effort often diverge (Anderson Gosselin & Gagné, 2011; Feuerstein, 1992), even though subjective ratings of effort correlate with performance. The lack of correlation between subjective and objective measures in these studies may be in part due to the two tasks using different mechanisms (Feuerstein, 1992) and/or individuals being unable to perceive increases in effort when task difficulty is increased (Anderson Gosselin & Gagné, 2011; Feuerstein, 1992). If the latter, then a robust objective or physiological measure of listening effort warrants development to remove subjectivity from the testing procedure to make it a clinically viable tool.

Understanding speech in adverse listening conditions is challenging, even for young adults with uncompromised hearing. Comprehending an incoming speech signal requires integrating phonemes, syllables, words and sentences in order to succeed in correctly identifying the intended message. Cognitive processes such as working memory may be recruited when listening to speech in noise in order to make verbal inferences, to segregate the speech signal from background noise, and to suppress distracting information. The current finding that individuals with lower working memory capacity do not differentiate effort ratings when spectral resolution is at its most challenging may have significant clinical implications, especially in older populations where hearing-impairment is commonplace and cognitive decline is known to co-occur. Future studies investigating the connection between listening effort, speech performance outcomes and working memory capacity will not only provide direction for increasing sensitivity of clinical assessments, but will also assist to refine models of cognitive hearing such as the Ease of Language Understanding model (Rönnberg, 2003; Rönnberg et al., 2013).

3.8 Acknowledgements

This research was supported by the HEARing CRC, established and supported by the Cooperative Research Centres Programme – Business Australia, and Macquarie University. We thank members of the Audiology and Hearing Research Group at Macquarie University, particularly Chi Yhun Lo, Louise Granger and Gabrielle Martinez for data collection, and members of the Linnaeus Centre HEAD, the Swedish Institute for Disability Research, Linköping University, especially Henrik Danielsson and Örjan Dahlström.

3.9 References

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47*(2), 53-71.

Anderson Gosselin, P., & Gagné, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*(3), 944-958.

Arlinger, S., Lunner, T., Lyxell, B., & Pichora-Fuller, K. (2009). The emergence of cognitive hearing science. *Scandinavian Journal of Psychology, 50*(5), 371-384.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual Review of Psychology, 63*, 1-29.

Bench, R., & Doyle, J. (1979). *The Bamford-Kowal-Bench/Australian version (BKB/A) Standard Sentence Lists.* Carlton, Victoria: Lincoln Institute.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America, 84*(1), 101-114.

Borg, G. A. (1982). Psychophysical bases of perceived exertion. *Medicine & Science in Sports & Exercise, 14*(5), 377-381.

Bregman, A. S. (1994). Auditory Scene Analysis: The Perceptual Organization of Sound. Cambridge, MA: MIT Press.

Conway, A. R., & Engle, R. W. (1994). Working memory and retrieval: a resource-dependent inhibition model. *Journal of Experimental Psychology: General, 123*(4), 354.

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: the importance of working memory capacity. *Psychonomic Bulletin & Review, 8*(2), 331-335.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General, 134*(2), 222.

Dillon, H. (1982). A quantitative examination of the sources of speech discrimination test score variability. *Ear and Hearing, 3*(2), 51-58.

Edwards, B. (2007). The future of hearing aid technology. *Trends in Amplification, 11*(1), 31-46.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: ease of listening, word recognition, and attentional effort. *Ear and Hearing, 13*(2), 80-86.

Flynn, M. C., & Dowell, R. C. (1999). Speech perception in a communicative context: an investigation using question/answer pairs. *Journal of Speech, Language, and Hearing Research, 42*(3), 540-552.

Foo, C., Rudner, M., Rönnberg, J., & Lunner, T. (2007). Recognition of speech in noise with new hearing instrument compression release settings requires explicit cognitive storage and processing capacity. *Journal of the American Academy of Audiology, 18*(7), 618-631.

Gisselgård, J., Petersson, K. M., & Ingvar, M. (2004). The irrelevant speech effect and working memory load. *Neuroimage, 22*(3), 1107-1116.

Grady, C. (2012). The cognitive neuroscience of ageing. *Nature Reviews Neuroscience, 13*(7), 491-505.

Grant, K. W., & Seitz, P. F. (2000). The recognition of isolated words and words in sentences: individual variability in the use of sentence context. *The Journal of the Acoustical Society of America, 107*(2), 1000-1011.

Hawkins, D. B., & Yacullo, W. S. (1984). Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation. *Journal of Speech and Hearing Disorders, 49*(3), 278-286.

Hua, H., Emilsson, M., Ellis, R., Widén, S., Möller, C., & Lyxell, B. (2014). Cognitive skills and the effect of noise on perceived effort in employees with aided hearing impairment and normal hearing. *Noise and Health, 16*(69), 79.

Hua, H., Anderzén-Carlsson, A., Widén, S., Möller, C., & Lyxell, B. (2015). Conceptions of working life among employees with mild-moderate aided hearing impairment: a phenomenographic study. *International Journal of Audiology, 54*(11), 1-8.

Humphreys, M. S., & Revelle, W. (1984). Personality, motivation, and performance: a theory of the relationship between individual differences and information processing. *Psychological Review, 91*(2), 153.

Keidser, G., Dillon, H., Mejia, J., & Nguyen, C.-V. (2013). An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function. *International Journal of Audiology, 52*(11), 795-800.

Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *International Journal of Audiology, 45*(9), 503-512.

Lin, F. R., & Ferrucci, L. (2012). Hearing loss and falls among older adults in the United States. *Archives of Internal Medicine, 172*(4), 369-371.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *International Journal of Audiology, 42*(1), 49-S58.

Lunner, T., Rudner, M., & Rönnberg, J. (2009). Cognition and hearing aids. *Scandinavian Journal of Psychology, 50*(5), 395-403.

Lyxell, B., Sahlen, B., Wass, M., Ibertsson, T., Larsby, B., Hällgren, M., & Mäki-Torkko, E. (2008). Cognitive development in children with cochlear implants: relations to reading and communication. *International Journal of Audiology, 47*(2), 47-S52.

Mehta, R. K., & Agnew, M. J. (2012). Influence of mental workload on muscle endurance, fatigue, and recovery during intermittent static work. *European Journal of Applied Physiology, 112*(8), 2891-2902.

Miles, C., Jones, D. M., & Madden, C. A. (1991). Locus of the irrelevant speech effect in short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*(3), 578.

Nachtegaal, J., Kuik, D. J., Anema, J. R., Goverts, S. T., Festen, J. M., & Kramer, S. E. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: results from an internet-based national survey on hearing. *International Journal of Audiology, 48*(10), 684-691.

Park, D. C., & Reuter-Lorenz, P. (2009). The adaptive brain: aging and neurocognitive scaffolding. *Annual Review of Psychology, 60*, 173-196.

Peters, M. L., Godaert, G. L., Ballieux, R. E., van Vliet, M., Willemsen, J. J., Sweep, F. C., & Heijnen, C. J. (1998). Cardiovascular and endocrine responses to experimental stress: effects of mental effort and controllability. *Psychoneuroendocrinology, 23*(1), 1-17.

Picard, M., Girard, S. A., Simard, M., Larocque, R., Leroux, T., & Turcotte, F. (2008). Association of work-related accidents with noise exposure in the workplace and noise-induced hearing loss based on the experience of some 240,000 person-years of observation. *Accident Analysis & Prevention, 40*(5), 1644-1652.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America, 97*(1), 593-608.

Pichora-Fuller, M. K. (2003). Cognitive aging and auditory information processing. *International Journal of Audiology, 42*(2), 26-32.

Pichora-Fuller, M. K. (2006). Perceptual effort and apparent cognitive decline: implications for audiologic rehabilitation. *Seminars in Hearing,* 27(1), 284-293.

Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: implications for hearing aid fitting and audiologic rehabilitation. *Trends in Amplification, 10*(1), 29-59.

Picou, E. M., & Ricketts, T. A. (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *International Journal of Audiology, 53*(6), 418-426.

Rabbitt, P. (1991). Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Oto-Laryngologica, 111*, 167-176.

Rönnberg, J., Arlinger, S., Lyxell, B., & Kinnefors, C. (1989). Visual evoked potentials relation to adult speechreading and cognitive function. *Journal of Speech, Language, and Hearing Research, 32*(4), 725-735.

Rönnberg, J. (2003). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: a framework and a model. *International Journal of Audiology, 42*(1), 68-76.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., . . . Pichora-Fuller, M. K. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience,* 7: 13.

Rosen, V. M., & Engle, R. W. (1998). Working memory capacity and suppression. *Journal of Memory and Language, 39*(3), 418-436.

Rudner, M., Foo, C., Rönnberg, J., & Lunner, T. (2009). Cognition and aided speech recognition in noise: Specific role for cognitive factors following nine-week experience with adjusted compression settings in hearing aids. *Scandinavian Journal of Psychology, 50*(5), 405-418.

Rudner, M., Ng, H. N., Rönnberg, N., Mishra, S., Rönnberg, J., Lunner, T., & Stenfelt, S. (2011). Cognitive spare capacity as a measure of listening effort. *Journal of Hearing Science, 1*(2), 47-49.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology, 23*(8), 577-589.

Salamé, P., & Baddeley, A. D. (1989). Effects of background music on phonological short-term memory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 41*(1-A), 107-122.

Salthouse, T. A. (2004). What and when of cognitive aging. *Current Directions in Psychological Science, 13*(4), 140-144.

Veneman, C. E., Gordon-Salant, S., Matthews, L. J., & Dubno, J. R. (2013). Age and measurement time-of-day effects on speech recognition in noise. *Ear and Hearing, 34*(3), 288.

Wayne, R. V., & Johnsrude, I. S. (2015). A review of causal mechanisms underlying the link between age-related hearing loss and cognitive decline. *Ageing Research Reviews, 23*, 154-166.

Wingfield, A. (1996). Cognitive factors in auditory performance: context, speed of processing, and constraints of memory. *Journal of the American Academy of Audiology, 7*(3), 175-182.

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing loss in older adulthood what it is and how it interacts with cognitive performance. *Current Directions in Psychological Science, 14*(3), 144-148.

World Health Organization. (2010). *International classification of functioning, disability and health: ICF*: World Health Organization.

Wouters, J., & Berghe, J. V. (2001). Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system. *Ear and Hearing, 22*(5), 420-430.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing, 32*(4), 498-510.

Chapter 4 Objective assessment of listening effort: co-registration of pupillometry and EEG

Special Issue: Australian Hearing Hub

Objective Assessment of Listening Effort: Coregistration of Pupillometry and EEG

Trends in Hearing
Volume 21: 1–13
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/2331216517706396
journals.sagepub.com/home/tia
SAGE

Kelly Miles[1,2,3], Catherine McMahon[1,2], Isabelle Boisvert[1,2], Ronny Ibrahim[1,2], Peter de Lissa[2,4], Petra Graham[5], and Björn Lyxell[3]

[1] Centre for Implementation of Hearing Research, Macquarie University, Sydney, Australia

[2] The HEARing Cooperative Research Centre, Melbourne, Australia

[3] Linnaeus Centre for HEaring And Deafness (HEAD), Swedish Institute for Disability Research, Linköping University, Sweden

[4] Department of Psychology, Australian Hearing Hub, Macquarie University, Sydney, Australia

[5] Department of Statistics, Australian Hearing Hub, Macquarie University, Sydney, Australia

Note that the following paper has been slightly amended after publication for consistency across the thesis. The analyses and results were not changed in any way.

4.1 Abstract

Listening to speech in noise is effortful, particularly for people with hearing impairment. While it is known that effort is related to a complex interplay between bottom-up and top-down processes, the cognitive and neurophysiological mechanisms contributing to effortful listening remain unknown. Therefore a reliable physiological measure to assess effort remains elusive. This study aimed to determine whether pupil dilation and alpha power change, two physiological measures suggested to index listening effort, assess similar processes. Listening effort was manipulated by parametrically varying spectral resolution (16- and 6-channel vocoding) and speech reception thresholds (SRT; 50% and 80%) while 19 young, normal-hearing adults performed a speech recognition task in noise. Results of offline sentence scoring showed discrepancies between the target SRTs and the true performance obtained during the speech recognition task. For example, in the SRT80% condition, participants scored an average of 64.7%. Participants' true performance levels were therefore used for subsequent statistical modelling. Results showed that both measures appeared to be sensitive to changes in spectral resolution (channel-vocoding), while pupil dilation only was also significantly related to their true performance levels (%) and task accuracy (i.e., whether the response was correctly or partially recalled). The two measures were not correlated, suggesting they each may reflect different cognitive processes involved in listening effort. This combination of findings contributes to a growing body of research aiming to develop a physiological measure of listening effort.

4.2 Introduction

Listening to speech in noise is a complex and effortful task, requiring a dynamic interplay between bottom-up and top-down processing (Arlinger, Lunner, Lyxell, & Pichora-Fuller, 2009; Rönnberg, Rudner, Foo, & Lunner, 2008; Zekveld, Heslenfeld, Festen, & Schoonhoven, 2006). Importantly, the effort required to listen to speech in noise is a commonly reported complaint in people with hearing impairment (Arlinger et al., 2009; Hawkins & Yacullo, 1984; Wouters & Berghe, 2001) that is not currently captured in standard clinical speech tests. Many factors may contribute to increased effort associated with listening to speech in noise, including age (Gosselin & Gagné, 2011; Larsby, Hällgren, Lyxell, & Arlinger, 2005; Tun, McCoy, & Wingfield, 2009) and cognitive influences such as working memory capacity and attention (Arlinger et al., 2009; Pichora-Fuller, 2006; Rudner, Lunner, Behrens, Thorén, & Rönnberg, 2012).

The term "listening effort" has been defined as "the mental exertion required to attend to, and understand, an auditory message" and has been studied from multiple perspectives (see McGarrigle et al. 2014, for a review). Adverse health effects of prolonged mental effort, particularly with an effort-reward imbalance (Kuper, Singh-Manoux, Siegrist, & Marmot, 2002; Siegrist, 1996), have been linked to fatigue (Mehta & Agnew, 2012), cardiovascular strain (Peters et al., 1998) and stress (Hua et al., 2014). For listening-related effort in particular, adults with hearing loss report increased incidence of fatigue (Hua, Anderzén-Carlsson, Widén, Möller, & Lyxell, 2015; Kramer, Kapteyn, & Houtgast, 2006; Pichora-Fuller, 2003), are absent from work more frequently (Kramer et al., 2006), take longer to recover after work (Nachtegaal et al., 2009), and may withdraw from society (Weinstein & Ventry, 1982). There is also evidence that children with hearing loss experience greater fatigue than their normal-hearing peers, in part due to the effort required to listen to their

teacher and interact with classmates, typically in acoustically poor classroom environments (Hornsby, Werfel, Camarata, & Bess, 2014).

Yet despite these negative health and social consequences of effortful listening, a reliable physiological measurement of listening effort remains elusive (Bernarding, Strauss, Hannemann, Seidler, & Corona-Strauss, 2013). Current speech perception assessments only provide a crude estimation of the limitations of hearing impairment, and do not typically consider the cognitive influences related to effort (Schneider, Pichora-Fuller, & Daneman, 2010; Wingfield, Tun, & McCoy, 2005) and the combination of, or interactions between, age and cognitive factors (Pichora-Fuller & Singh, 2006). Simultaneous evaluation of listening effort during speech recognition in noise could increase sensitivity of these assessments and guide device selection and settings as well as rehabilitation strategies.

A wide range of methods and tools have been used to explore listening effort that may better reflect the cognitive challenges which individuals with hearing loss face in real-world environments. Such measures have included subjective ratings (scales and questionnaires), dual tasks (performance measures on one task while the difficulty of a concurrent task varies) and physiological measures such as changes in brain oscillations, pupillometry, skin conductance and cortisol levels (see McGarrigle et al., 2014 for a comprehensive review). At present, pupil dilation and EEG are the most-cited physiological measures that have the clinical potential to assess listening effort due to their non-invasiveness, increasing portability and user-friendliness (Badcock et al., 2013; Mele & Federici, 2012), and ability to be used during standard clinical speech perception assessments.

Changes in the pupillary response, which is under the physiological control of the locus coeruleus-norepinepherine (LC-NE) system (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010) has been argued to reflect increased processing load (Beatty & Wagoner, 1977;

Granholm, Asarnow, Sarkin, & Dykes, 1996). Pupil size has been shown to be larger during sentence encoding when performance levels are low (e.g., 50% Speech Reception Threshold (SRT)) in comparison to higher performance levels (e.g., 84% SRT; Koelewijn, Zekveld, Festen, Rönnberg, & Kramer, 2012; Kramer, Kapteyn, Festen, & Kuik, 1997; Zekveld, Kramer, & Festen, 2010; Zekveld, Festen, & Kramer, 2014; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; Zekveld & Kramer, 2014). In addition to varying signal-to-noise ratios (SNRs) and SRTs, other speech stimuli manipulations have been shown to affect pupil dilation. For example, pupil size is larger when listening to speech in single-talker maskers than in fluctuating noise (Koelewijn et al., 2014; Koelewijn, Zekveld, Festen, Rönnberg, et al., 2012) or stationary noise (Koelewijn, Zekveld, Festen, & Kramer, 2012), or when the masking speech is the same gender as the speech signal (Zekveld, Rudner, Kramer, Lyzenga, & Rönnberg, 2014). Changing the complexity of the task through linguistic manipulation such as lexical competition (Kuchinsky et al., 2013) and sentence difficulty also increases the pupil size in the more linguistically challenging conditions (Piquado, Isaacowitz, & Wingfield, 2010; Wendt, Dau, & Hjortkjær, 2016). Degrading the spectral resolution of a speech stimulus through channel-vocoding also increases pupil size (Winn, Edwards, & Litovsky, 2015). Collectively, these studies suggest that the pupillary response changes with task difficulty, which may reflect the increased effort associated with more challenging tasks.

Changes in brain oscillations is another physiological measure that has shown systematic changes associated with a wide variety of cognitive processes (Başar, Başar-Eroglu, Karakaş, & Schürmann, 2001; Herrmann, Fründ, & Lenz, 2010; Klimesch, 1996, 1999, 2012; Ward, 2003). For example, enhancement in the alpha frequency band (8-12 Hz) has been observed using working memory tasks with various types of stimuli, including syllables (Leiberg, Lutzenberger, & Kaiser, 2006) and single words (Karrasch, Laine, Rapinoja, & Krause, 2004; Pesonen, Björnberg, Hämäläinen, & Krause, 2006). Obleser and colleagues (2012) proposed

that acoustic degradation and working memory load may similarly affect alpha oscillations due to the greater allocation of working memory resources required to comprehend an acoustically degraded signal (Piquado et al., 2010; Rabbitt, 1968; Wingfield et al., 2005). When memory load and spectral resolution were most challenging, they demonstrated that alpha enhancement was superadditive, suggesting that the same alpha network might index both. Similar findings have been replicated using different sources of acoustic degradation during a digit or word comprehension task (Petersen, Wöstmann, Obleser, Stenfelt, & Lunner, 2015; Wöstmann, Herrmann, Wilsch, & Obleser, 2015; see Strauß, Wöstmann, & Obleser, 2014, for review).

In our previous study (McMahon et al., 2016), we examined the changes in alpha power and pupil dilation during a speech recognition in noise task using channel-vocoded sentences (16- and 6-channel) and a 4-talker babble-noise varying from -7 dB to +7 dB SNR. This demonstrated that the change in alpha power significantly declined with increasing SNRs for 16-channel vocoded sentences, but remained relatively unchanged for 6-channel sentences. Pupil dilation showed a similar negative linear correlation for 16-channel vocoded sentences, which also was not observed for 6-channel vocoded sentences (instead showing a strong cubic relationship with SNR). Finally, changes in pupil dilation and alpha power were not correlated. However, differences in performance may explain the lack of correlation, as fixed SNRs were used across all participants. As such, for the present study, performance levels were fixed using individually obtained 50 and 80% SRTs for both 16-channel and 6-channel vocoded sentences. The signal-to-noise ratio was modulated to achieve these performance levels (and were therefore different across the participants). Manipulating two types of acoustic degradation approximately simulates listening in noise with a cochlear implant (Friesen, Shannon, Baskent & Wing, 2001) and is relevant if physiological measures of listening effort are to be applied in a clinical setting.

Alpha power and pupil dilation have both been proposed to be physiological measures of listening effort with the possibility of implementation into clinical practice. Gaining a better understanding of how these physiological measures respond to changes in task difficulty within a population with normal hearing and cognition is important in order to interpret its behaviour within an older population with hearing loss. The current study aimed to determine how pupil dilation and alpha oscillations change when parametrically varying both spectral resolution (using 16-channel and 6-channel vocoding) and speech recognition performance. Simultaneous recordings of pupil dilation and alpha oscillations will determine whether the measures are associated and might enable insight into whether they index the same aspect of listening effort.

4.3 Materials and methods

Participants

Twenty-seven normal-hearing monolingual English-speaking participants were recruited for the study. Two participants did not attend all testing sessions and were therefore excluded. As one of the main aims of this study was to assess whether EEG and pupil dilation correlated with each other, only participants who had 65% accepted trials in both measures were included (n=19). Participants (12 women, 7 men) had a mean age of 27 years (SD = 4.28, range = 22-34 years). All had distortion-product otoacoustic emissions between 1 – 4 kHz, consistent with typical hearing or a mild sensorineural loss only and were right-handed as assessed by *The Assessment and Analysis of Handedness: the Edinburgh Inventory* (Oldfield, 1971).

Speech materials and background noise

Bamford-Kowal-Bench sentences adapted for Australian-English (Bench & Doyle, 1979) were recorded by a native Australian-English female speaker. The sentences and background noise (4-talker babble) were vocoded using custom MATLAB scripts where the frequency

range 50-6000 Hz was divided into 6 or 16 logarithmically spaced channels. The amplitude envelope was then extracted by taking the absolute value from the Hilbert transform from each channel. The extracted envelope was used to modulate noise with the same frequency band. Each band of noise was then recombined to produce the channel-vocoded sentences and background noise. The root mean square (RMS) levels of the sentences and background noise were equalized in MATLAB after vocoding.

Speech reception thresholds

Automated adaptive speech-in-noise software developed by the National Acoustic Laboratories was used to obtain SRTs (see Keidser, Dillon, Mejia & Nguyen, 2013, for a comprehensive review of the algorithm). The adaptive test has been validated with similar speech materials to the current study, with participants with normal hearing (n=12) and hearing loss (n=63), showing a standard deviation of 1.27 dB and 1.24 dB, respectively (Keidser et al., 2013). This suggests that the test results are reliable. The BKB-A sentences were presented at 65 dB SPL, and the background noise was adaptively adjusted to obtain each of the SRTs. The adaptive procedure consisted of three phases. *Phase 1*: 5 dB steps until 4 sentences were completed, including one reversal, *Phase 2:* 2 dB steps until a minimum of 4 sentences were completed, and the phase's standard error (SE) was 1 dB or below, and *Phase 3:* 1 dB steps until 16 sentences (from the end of phase 2) were completed, with a SE of 0.80 or below, or the maximum number of 32 sentences was reached (note that the minimum was 16 sentences). When the SE reached 0.80 or below, the test terminated and recorded the SNR. This procedure was conducted across both 50% and 80% SRTs.

A sound-attenuated room was used during the testing sessions, and the equipment was calibrated prior to each participant's arrival. The speaker was positioned at one meter and zero degrees azimuth from the participant. Participants were informed they would hear sentences in noise and were instructed to repeat back all of the words of the sentence they

heard. Two SRTs (50%, 80%) for two vocoding conditions (16-channel, 6-channel) were collected, resulting in four conditions. Table 1 shows the mean and standard deviations of the SNRs for each SRT.

Table 1. Means and standard deviations of SNR (dB), true performance obtained during the physiological testing session (%), and pearson's r correlation coefficients of pupil dilation and alpha power, presented by SRT and channel-vocoding (first-two columns), and channel-vocoding (collapsed across SRT) and SRT (collapsed across channel-vocoding) in the last two columns. SNR = signal-to-noise ratios; SRT = speech reception thresholds.

SRT %	*50*		*80*		-	-	*50*	*80*
Channel vocoding	*6*	*16*	*6*	*16*	*6*	*16*	-	-
SNR dB	0.4 (1.7)	-1.7 (1.4)	3.7 (1.9)	0.7 (1.7)	2.1 (2.4)	-0.7 (1.9)	-0.6 (2.0)	2.0 (2.4)
Performance %	39.2 (16.5)	53.0 (15.8)	60.6 (15.0)	68.8 (13.5)	49.9 (19.0)	60.9 (16.6)	46.1 (17.4)	64.7 (14.7)
Pearson's r	0.02 (0.14)	0.08 (0.20)	-0.04 (0.15)	-0.01 (0.17)	-0.01 (0.09)	0.03 (0.11)	0.06 (0.13)	-0.04 (0.08)

Physiological measures

The pupil and EEG recordings were measured simultaneously during the speech recognition task in a sound-attenuated and magnetically shielded room. Each participant was assigned randomly to Block A or Block B. Each block contained 220 sentences divided into the 4 conditions (6ch50%SRT, 6ch80%SRT, 16ch50%SRT, and 16ch80%SRT) in which the sentences were swapped in each block. For example, in Block A, a sentence presented in the 6ch50%SRT condition was presented in the 16ch80%SRT condition in Block B. Sentences were randomized during presentation. Each participant's SNR obtained during the behavioral session was used to present the sentences in each of the conditions. The top panel of Figure 1 shows the presentation protocol. Participants were instructed to repeat the sentence at the offset of the noise. Responses were recorded using a voice recorder and video-camera set up directly in front of them to capture their face during recording allowing more accurate scoring of their responses at a later time. The sentences were scored at the word level (using the

standard BKB/A scoring criteria) by a native Australian-English speaker and the percentage correct was averaged for each condition.

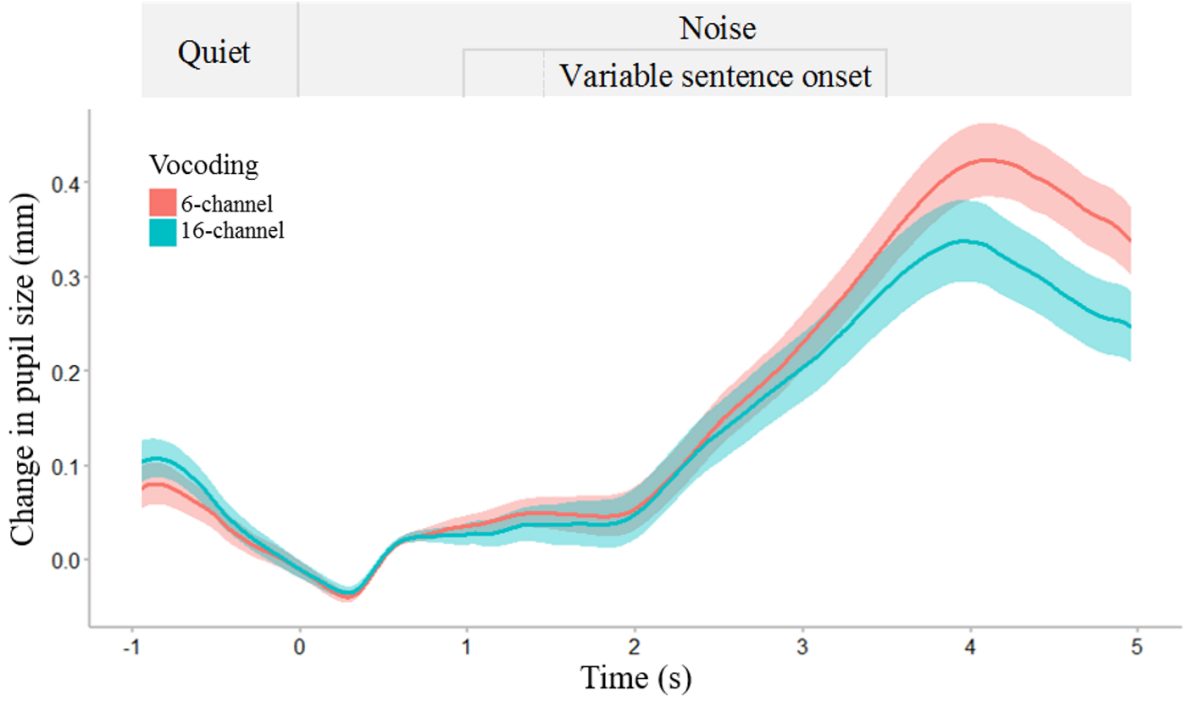

Figure 1. Average pupil size over time for all trials and participants, for 16- and 6-channel vocoding. The 0 s time-point refers to the beginning of noise. All sentences finished at the 3.5 s time-point. Shading represents ±1SE of the mean. The top panel represents the presentation protocol.

Pupil recording

The right pupil was recorded using an SR Research Eyelink 1000 tower mount system at a 1000 Hz sampling rate. Stimuli were presented through Experiment Builder software v1.10.1241. Prior to the task, the equipment was calibrated using a 9-point calibration grid on the screen. Pupil activity was recorded continuously until the experiment terminated. Offline, single-trials were processed with Dataviewer software (version 1.11.1), and compiled into single-trial pupil-diameter waveforms (-1 s to 5 s) for further processing and analyses using customized MATLAB scripts. Blink identification was determined on a trial-by-trial basis as pupil sample sizes smaller than three standard deviations below the mean pupil diameter.

Trials containing more than 15% of the trial samples detected in blink were rejected. In the remaining accepted trials, linear interpolation was used to reduce lost samples and artefacts from blinks (Siegle, Ichikawa, & Steinhauer, 2008; Zekveld, et al., 2010). Samples affected by blinks were interpolated from 66 ms preceding the onset of a blink to 132 ms following the offset of a blink. Data were smoothed using a 5-point moving average. Trials were then averaged across conditions for each participant. Regions of interest included baseline in noise (0-1 s) and the encoding period (2-6 s).

For each trial, relative percent change was calculated as maximum pupil size during encoding minus mean pupil size during baseline in noise, divided by the mean baseline in noise. This was then multiplied by 100 in order to report percent change from baseline. See Figure 1 for an example of the pupil response (shown in millimeters, not percent change), during the experiment.

EEG recording

The continuous EEG was recorded with a 32-channel SynampsII Neuroscan amplifier. Thirty electrodes were positioned on the scalp in a standard 10-20 configuration (FP1 and FP2 were disabled as the participants rested their foreheads on the eyetracker support). The ground electrode was located between Fz and FPz electrodes. Electrical activity was recorded from both mastoids, with M1 set as the online reference. Ocular movement was recorded with bipolar electrodes placed at the outer canthi, and above and below the left eye. Electrode impedances were kept below 5 kΩ. The signal from the scalp electrodes was sampled at 1000 Hz, band-pass filtered between .01 and 100 Hz, and notch filtered online at 50 Hz.

Ocular artefact rejection was performed using Neuroscan software using a standard ocular reduction algorithm. Post-processing was conducted in Fieldtrip-MATLAB (Oostenveld et al., 2011). The EEG data were epoched from -1 second to 5 seconds avoiding stimulus boundary artefacts caused by the filtering process. A two-pass reversal Butterworth filter with

cut-off frequencies of 0.5 Hz- 45 Hz was applied to remove any drifts and high frequency noise that might occur. Band-pass filtering was used instead of high-pass filtering as in Obleser et al., (2012). Trials containing a variance exceeding 300 μV^2 were removed from further analyses. Trials were then two-pass band-pass filtered between 8-12 Hz to extract the alpha oscillation (the mean absolute value of the filtered time series). The absolute value of the alpha band was extracted from the parietal electrodes (P3, P4, and Pz) during the encoding period (one second duration finishing 200 ms before the end of the sentence) and baseline in noise (300 ms-800 ms after the noise onset) on a trial by trial basis. See Figure 2 for a time-frequency representation of the EEG activity.

For each trial, relative percent change was calculated as mean alpha power during encoding minus mean alpha power during baseline, divided by mean alpha power during baseline. This was then multiplied by 100 in order to report percent change from baseline.

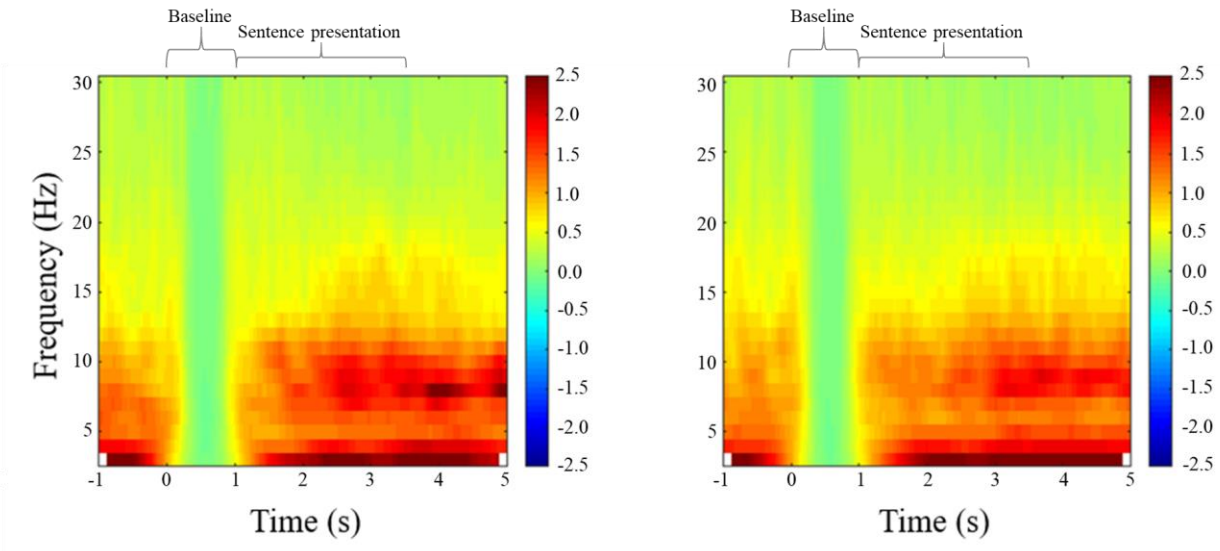

Figure 2. Time-frequency representation of the EEG activity averaged across all participants, in the parietal region, for 16- and 6-channel vocoding. The time-frequency representations are relative to the activity occurring during the 1 s of noise beginning at the 0 s time-point

(baseline). All sentences finished at the 3.5 s time-point. The colors represent frequency power levels.

4.4 Statistical methods

As the range of absolute values was considerably different between pupil dilation and alpha power change, relative percent change from baseline was used for all statistical analyses in order to facilitate comparisons between the two measures.

All analyses were performed in R version 3.2.1 using the nlme package (Pinheiro, Bates, DebRoy, & Sarkar, 2014). Linear mixed-effects (LME) models with a random intercept for individual were used for all analyses to control for repeated measures over different levels of the factors on individuals. Models for pupil size and alpha power were first assessed by including interactions. If there was no significant interaction, the main effects model was presented. P-values less than 0.05 were considered significant for all analyses.

Four LME regression models were developed to determine the effects of the task parameters on pupil dilation and alpha power (see Table 2). The first model assessed the effect of SRT (50% and 80%) and vocoding (16-channel and 6-channel) on these measures. Variability in performance for the targeted SRTs existed within the physiological experiment, despite using a validated adaptive method in the behavioral session to obtain these SRTs. This was particularly evident for the 80% SRT, where participants' true performance on the sentence recall task ranged from an average of 36% to 93% (see Table 4 *True performance %* for mean performance obtained for all conditions). To account for this variability, a second LME regression model was introduced, with true performance levels (i.e., individual speech recognition scores) and vocoding as predictor variables. A third LME model was developed to assess whether task accuracy influenced the measures. This was done because the reason for an incorrect or not recalled response is variable and generally unknown. For example, if a

participant is inattentive during a trial and does not recall a sentence, the physiological response may be different compared to if the participant invested effort to hear a sentence that was too challenging to perceive. Removing incorrect or not recalled sentences may reduce variability in the measures. Therefore, only correct or partially recalled sentences were analyzed in the third model. Finally, a fourth model was constructed to assess whether changes in pupil size and alpha power were not merely due to changes in SNR (i.e., varying loudness across the conditions). Individual SNRs for each participant and condition were used in the model.

To account for repeated measures on individuals, correlations presented in the results section are the average of the correlation coefficients calculated for each participant (average of the 19 coefficients for each of the 8 conditions). As there were unequal trials across the pupil and alpha measures, only those trials that were accepted in both measures were used in the correlation analyses (n = 3019).

4.5 Results

SRT and vocoding

To determine the effect of SRT and vocoding on changes in pupil dilation or alpha power, LME regression models with SRT and vocoding as predictor variables were developed. For pupil dilation, there was no significant interaction term ($p = 0.12$). A main effects model indicated no effect of SRT ($p = 0.91$), and a significant effect of vocoding ($p < 0.01$) on pupil size, which was 1.49% larger in the 6-channel condition compared to the 16-channel condition. For alpha power, there was no significant interaction term ($p = 0.62$). A main effects model indicated no effect of SRT ($p = 0.29$), and a significant effect of vocoding ($p = 0.03$). Alpha power change was -30.0% lower in the 6-channel condition compared to the 16-

channel condition. Figure 3 shows the mean of maximum pupil size and alpha power change relative to baseline.

Table 2. Results for the Linear Mixed-Effects Models. Reference levels: 50% SRT, 16-channel vocoding, correct recall (task accuracy). True performance (i.e., the actual percentage of speech recognized during the physiological testing session) was modelled in addition to SRT, as off-line sentence scoring was shown to deviate from target SRTs. SRT = speech reception thresholds; CI = confidence interval; SNR = signal-to-noise ratios.

Model	**Pupil size**			**Alpha power**		
	Estimate	95% CI	*p* value	Estimate	95% CI	*p* value
Relative change in pupil/alpha						
Intercept	12.156	[9.145, 15.077]	<0.001	150.727	[84.345, 187.475]	<0.001
SRT	0.045	[-0.773, 0.863]	0.915	-14.818	[-12.472, 42.108]	0.287
Vocoding	1.491	[0.673, 2.308]	<0.001	-29.991	[-57.285, -2.696]	0.031
Relative change in pupil/alpha						
Intercept	14.908	[11.544, 18.271]	<0.001	134.693	[62.225, 207.160]	0.003
Performance level	-0.046	[-0.072, -0.019]	<0.001	0.14	[-0.722, 1.001]	0.751
Vocoding	0.986	[0.119, 1.852]	0.026	-28.261	[-57.310, 0.788]	0.057
Relative change in pupil/alpha						
Intercept	11.191	[8.242, 14.138]	<0.001	133.096	[81.166, 185.026]	<0.001
Task accuracy	1.612	[0.676, 2.549]	<0.001	7.661	[-23.206, 38.528]	0.626
Vocoding	1.571	[0.644, 2.499]	<0.001	-17.833	[-48.336, 12.670]	0.252
Relative change in pupil/alpha						
Intercept	12.864	[9.956, 15.774]	<0.001	130.475	[82.717, 178.233]	<0.001
SNR	0.018	[-0.153, 0.189]	0.836	-3.46	[-9.112, 2.191]	0.23

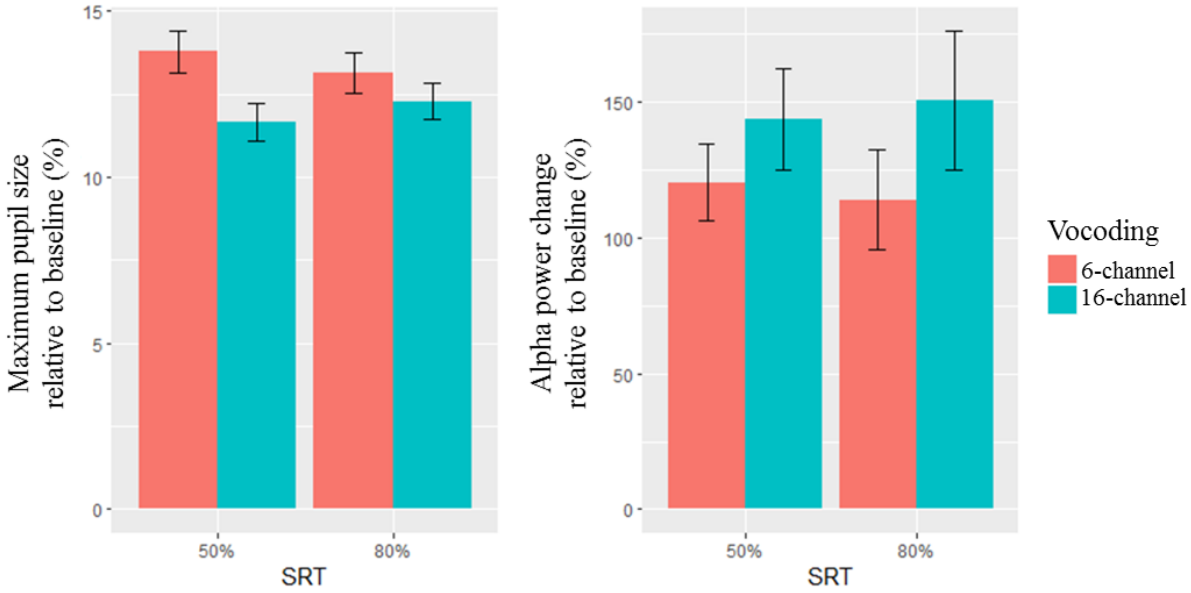

Figure 3. Mean ±1SE of maximum pupil size and alpha power change relative to baseline, by SRT and channel-vocoding.

True performance and vocoding

As previously discussed, the true performance obtained on the speech recognition task during the physiological testing session greatly differed from the targeted 50% and 80% SRTs (see Table 4). To determine whether this had any bearing on the physiological measures, LME regressions models with true performance (%) and vocoding as predictor variables were developed.

For pupil dilation, there was no significant interaction term ($p = 0.88$). A main effects model indicated a significant effect of performance level ($p < 0.01$), and vocoding ($p = 0.03$). As performance levels increased (towards 100%), the pupil size decreased by -0.05%. For alpha power, a LME regression model with performance level and vocoding as predictor variables indicated there was no significant interaction term ($p = 0.88$). A main effects model indicated no effect of performance level ($p = 0.75$). The inclusion of the participants' true performance level caused a loss of significance of channel-vocoding ($p = 0.06$) in alpha power compared to the SRT and vocoding model. A likelihood ratio test suggested no benefit of including

performance levels in the model (performance level and vocoding: log likelihood -24869.65 versus vocoding alone: log likelihood -24869.79). Removing performance level indicated that vocoding was significant ($p = 0.03$). Alpha power was 29.9% greater in the 16-channel condition compared to 6-channel.

Task accuracy and vocoding

A LME model including task accuracy (partially correct versus correct sentence recall) and channel-vocoding was developed. For pupil dilation, there was no significant interaction term ($p = 0.38$). A main effects model showed a significant effect of task accuracy ($p<0.01$) and a significant effect of vocoding ($p<0.01$). Pupil size for partially correct sentence recall was 1.61% larger than correctly recalled sentences and 1.57% larger for 6-channel vocoded sentences compared to 16-channel. For alpha power, there was no significant interaction term ($p = 0.17$). A main effects model showed no effect of task accuracy ($p = 0.63$) and no effect of vocoding ($p = 0.25$).

SNR

Finally, to determine whether SNR was influencing pupil size or alpha power, SNR was entered into an LME as a predictor variable, however this showed no significant effect on pupil size ($p = 0.84$) or alpha power ($p = 0.23$).

Correlation between pupil size and alpha power change

At the individual level, pupil size was not significantly correlated with alpha power change for any of the correlations, (n = 19, $p>0.05$ for all correlations; see table 1 for means and SDs), including collapsing across all conditions (n = 19, mean $r = 0.01$ $SD = 0.08$, $p>0.05$). Figure 4 shows individual Pearson's r coefficients for participants, by vocoding.

To assess whether the lack of correlation between the measures may be due to intra-individual variability, intra-class correlations (ICC) were conducted. ICC estimates and their 95% confidence intervals were calculated using the ICC package in R (Wolak, Fairbairn & Paulsen, 2012). In order to balance the dataset for ICC, the minimum number of trials available per subject, per condition, was first determined (n = 19). The results of the ICC analysis shows a weak-to-strong degree of reliability for the alpha measurements (Table 2), and a very high degree of intra-individual reliability for the pupil measurements (Table 3). Alpha power was more variable than the pupil dilation measure, however, even in the conditions where there was strong reliability across both measures (e.g., 16-channel) the two measures were not correlated (Table 1).

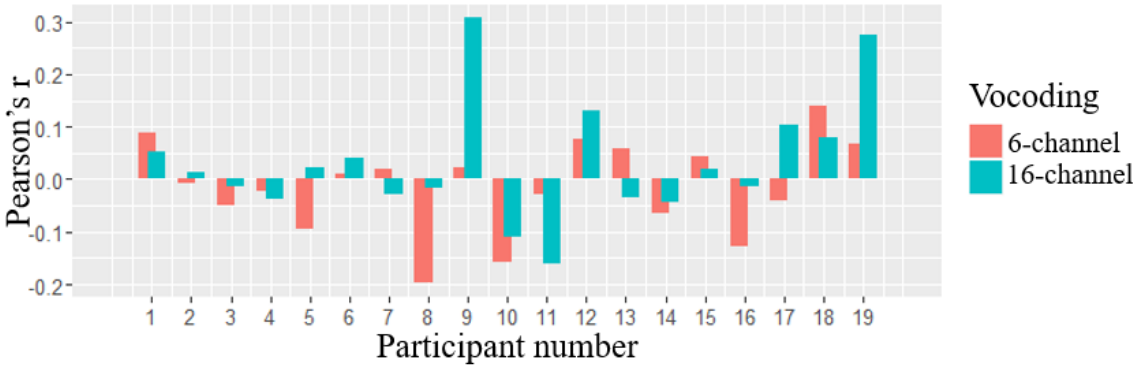

Figure 4. Pearson's r correlation coefficients of pupil dilation and alpha power by participant and channel-vocoding.

Table 3. Intraclass Correlation Coefficients (ICC) assessing intraindividual reliability for alpha power. SRT = speech reception thresholds; CI = confidence intervals.

SRT %	*50*		*80*		-	-	*50*	*80*
Channel vocoding	*6*	*16*	*6*	*16*	*6*	*16*	-	-
ICC	0.38	0.63	0.22	0.49	0.5	0.76	0.73	0.63
p value	0.055	<0.001	0.198	0.012	0.008	<0.001	<0.001	<0.001
95% CI	-0.11, 0.72	0.34, 0.83	-0.40, 0.65	0.08, 0.77	0.12, 0.77	0.58, 0.89	0.53, 0.88	0.35, 0.83

Table 4. Intraclass Correlation Coefficients (ICC) assessing intraindividual reliability for pupil dilation. SRT = speech reception thresholds; CI = confidence intervals.

SRT %	*50*		*80*		-	-	*50*	*80*
Channel vocoding	*6*	*16*	*6*	*16*	*6*	*16*	-	-
ICC	0.82	0.84	0.85	0.85	0.91	0.92	0.9	0.92
p value	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
95% CI	0.68, 0.92	0.71, 0.93	0.73, 0.93	0.73, 0.93	0.85, 0.96	0.86, 0.96	0.82, 0.95	0.85, 0.96

4.6 Discussion

The aim of this study was to examine the effects of increasing listening effort during a sentence recognition-in-noise task on pupil dilation and alpha power change. The second aim was to determine whether these physiological measures that were recorded simultaneously, were correlated. This would suggest that each respond to the same aspect of listening effort which may, itself, comprise multiple components (see Pichora-Fuller et al., 2016 for review). Listening effort was manipulated by parametrically varying the spectral content of the signal using channel-vocoding (16- and 6-channels) and performance (50% and 80% SRT). Specified SRTs were chosen rather than a fixed SNR to account for cognitive differences within the participant population (Souza & Arehart, 2015) and to investigate whether the large variability in alpha power change and pupil dilation change across the participants as reported by McMahon et al. (2016) was due to the influence of performance.

In the SRT and vocoding model, the more spectrally degraded 6-channel vocoded sentences elicited greater pupil dilation, on average, compared to the 16-channel sentences. This finding is consistent with previous studies where decreasing the spectral resolution of the signal (i.e., decreasing the number of channels in vocoded speech) systematically increases pupil diameter (Winn et al., 2015) suggesting listening was more effortful.

On average, alpha power was greater in the less spectrally degraded 16-channel condition, consistent with McMahon et al. (2016) which used the same channel-vocoded sentences and 4-talker babble background noise. This finding diverges from studies using less complex linguistic stimuli which have shown that decreased acoustic quality enhances alpha power (digit task: Obleser et al., 2012, Wöstmann et al., 2015; word comprehension: Becker, Pefkou, Michel, & Hervais-Adelman, 2013; Obleser & Weisz, 2012). This could suggest that spectrally degraded sentences influence the alpha network differently due to the increased linguistic complexity. In the easier 16-channel condition, because of the better spectral resolution, there may be less dependency on the semantic context to recognize a sentence, whereas in the more spectrally degraded 6-channel condition, there is a greater need to rely on semantic context to fill in the gaps. However, before semantic processing is engaged, at least some phonemes must be recognized in the incoming speech signal. This lower level acoustic/phonemic processing may be more demanding in the 6-channel condition, perhaps decreasing the possibility for more automated semantic recognition. Greater alpha power in the 16-channel condition may therefore reflect task-irrelevant inhibition due to the automaticity of sentence processing when the signal was clearer, while reduced alpha power in the 6-channel condition may reflect ongoing active processing (Klimesch, 2012; Weisz, Hartmann, Müller, & Obleser, 2011) due to the poorer spectral resolution.

It was anticipated that the pupil would be sensitive to changes in SRT, with greater increases in pupil size expected in the more cognitively demanding 50% SRT compared to 80% SRT condition. Contrary to expectations, however, this was not the case. During the physiological session, the true performance levels obtained for each SRT varied substantially from the target SRTs of 50% and 80%. On average, the true performance difference was only 18.6% between conditions, compared to the target SRTs which differed by 30% (i.e., the performance difference between 50% and 80% SRT). It is therefore possible that the

narrower range of the true performance contributed to the non-significant change in pupil dilation when modelling SRT.

Using SRTs is common practice to assess speech recognition in noise both clinically, and for research purposes (Best, Keidser, Buchholz, & Freeston, 2015; Lunner, 2003; Smits & Festen, 2013). Numerous pupillometry studies have varied task difficulty by using an adaptive method to reach a desired SRT while recoding the pupil response. In the current study, each individual's SNR was fixed during the physiological session in order to randomize stimuli presentation, as per Zekveld, Heslenfeld, Johnsrude, Versfeld, and Kramer (2014). To do this, participants' SNRs for 50% and 80% SRT were obtained in a prior behavioral session (see methods section). The adaptive test has been validated with similar speech materials for participants with normal hearing and hearing loss and showed less variability than the current study (Keidser et al., 2013). The higher variability in the current study may have resulted from acoustically degrading the sentence materials (i.e., channel-vocoding), or the different application of the SRTs across test sessions. Further, given that the behavioral session was considerably shorter in duration than the physiological session, the discrepancy between true performance and the target SRTs across the two sessions may be influenced by the participants' varying levels of motivation and fatigue between sessions. Irrespective of the cause of the variability, true performance levels were used in an alternative LME regression model, instead of the 50% and 80% SRTs.

True performance levels had no effect on alpha power, further suggesting that potential between-subjects variability in performance when sentences are presented in fixed SNRs (McMahon et al., 2016) is not likely to be influencing the change in alpha power. In line with previous studies (Kramer et al., 1997; Zekveld et al., 2010), increasing performance significantly decreased pupil size. Moreover, consistent with Winn et al. (2015), decreasing the spectral quality of the signal elicited greater pupil dilation, on average, even when

accounting for individuals' true performance, demonstrating that performance alone does not fully capture the effort required to process a spectrally degraded signal. This is particularly evident in studies showing that even when task accuracy is high (~100%), continuing to degrade the spectral resolution of the signal increases effort, as reflected in greater pupil size (Winn et al., 2015).

To further assess the relationship between the two physiological measures and task accuracy, only partially and correctly recalled sentences were examined. As incorrect sentence recall (or an absent response) could be due to many factors, including attention and/ or misperception (Kuchinsky et al., 2013) they were excluded from the analysis. Pupil size was significantly larger in the 6-channel condition, and also for partially recalled sentences, however there was no significant interaction between the two effects. Therefore, while decreased spectral resolution and partially recalled sentences appear to increase listening effort, when sentences are accurately recalled (correct response), there is no difference in listening effort between vocoding conditions, as indexed by pupil size. Unlike the pupil response, there was no difference in alpha power across the levels of task accuracy. This is in line with Obleser and colleagues (2012) who found task accuracy and alpha power were not correlated, although task accuracy was relatively high in their study (91-100%) and responses were either correct or incorrect. Removing the incorrect responses in the current study (n = 888), revealed that channel-vocoding no longer appeared to influence alpha power. However, this is possibly due to the decreased statistical power when removing incorrectly recalled sentences.

Consistent with previous studies examining speech recognition in noise, pupil size and alpha power were not significantly correlated within individuals (McMahon et al., 2016). This lack of correlation has been similarly reported in the reading domain (Scharinger, Kammerer, & Gerjets, 2015). As speculated by McMahon et al. (2016) this may be due to attention

mechanisms, such as individuals using different encoding and modifying strategies (c.f., Power and Petersen, 2013), the measures themselves being under the control of different attentional networks (c.f. Corbetta, Patel, & Shulman, 2008, for review of the different attention networks), or that each are encoding different aspects of listening effort.

Population parameters such as age, hearing, and cognitive ability, may interact with the pupil and alpha response, which may explain some of the variability which exists in the current literature. For example, Petersen et al. (2015) found that alpha power breaks down in participants with moderate hearing loss when spectral resolution and working memory capacity is at its most challenging level. Zekveld et al., (2011) have also demonstrated that the relationship between a smaller pupil response (which would reflect a decreased cognitive load) and increasing speech intelligibility was indeed weaker in people with hearing impairment. The pupil and alpha response when hearing is impeded either by internal acoustic degradation (such as a hearing loss) versus external acoustic degradation (such as presenting a degraded signal to individuals with/ without hearing impairment) may therefore not be entirely comparable.

Limitations of the current study include confining the analyses of alpha power change to the parietal area. While majority of the studies outlined in this paper found enhanced alpha activity in this location, it may be too restrictive given the added complexity of processing spectrally degraded sentences in noise. Furthermore, alpha band (8-12 Hz) may be too coarse, averaging two (or more) parallel processes. For example, 8-10 Hz is suggested to relate to attentional processes while 10-12 Hz may be related to linguistic-type activity such as semantic processing (cf. Klimesch, 1999, 2012, for review), where each selectively synchronize or desynchronize based on the stimulus. Future studies may wish to consider whole-head analysis and narrow-band frequencies which may provide further insight into different aspects of listening effort.

4.7 Conclusion

Both changes in pupil dilation and alpha power have been suggested to index listening effort. Understanding how these measures behave, and interact, during simultaneous measurement provides insight into what aspects of listening effort they may each be encoding. This will go a long way towards disentangling the multifaceted nature of listening effort, and will improve the prospect of using them to complement standard hearing assessments. Here, we began to address these issues in a young, normal-hearing population to better understand how they operate when sensory input is not compromised by hearing impairment, and cognition is robust. Both measures appeared to be sensitive to changes in spectral resolution (channel-vocoding), while pupil dilation provided further information about performance levels and task accuracy. Further, the measures were not correlated, suggesting they may be sensitive, or respond differently to, the different aspects of 'mental exertion' that comprise listening effort.

4.8 Acknowledgments

This study was supported by the Macquarie University Research Excellence Scheme, and the HEARing Cooperative Research Centre, established and supported under the Business Cooperative Research Centres Programme of the Australian Government. The authors thank Jorg Buchholz for assisting with protocol development, Nille Elise Kepp for help preparing the test materials, and Jaime Undurraga and Keiran Thompson for programming advice. The Departments of Linguistics, Psychology, and Statistics, are part of the Australian Hearing Hub, an initiative of Macquarie University, that brings together Australia's leading hearing and healthcare organizations to collaborate on research projects.

4.9 References

Arlinger, S., Lunner, T., Lyxell, B., & Pichora-Fuller, K. M. (2009). The emergence of cognitive hearing science. *Scandinavian Journal of Psychology, 50*(5), 371-384.

Badcock, N. A., Mousikou, P., Mahajan, Y., de Lissa, P., Thie, J., & McArthur, G. (2013). Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. *PeerJ, 1*, e38.

Başar, E., Başar-Eroglu, C., Karakaş, S., & Schürmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International Journal of Psychophysiology, 39*(2), 241-248.

Becker, R., Pefkou, M., Michel, C. M., & Hervais-Adelman, A. G. (2013). Left temporal alpha-band activity reflects single word intelligibility. *Frontiers in Systems Neuroscience*, 7: 121.

Bench, R., & Doyle, J. (1979). *The Bamford-Kowal-Bench/Australian version (BKB/A) Standard Sentence Lists.* Carlton, Victoria: Lincoln Institute.

Bernarding, C., Strauss, D. J., Hannemann, R., Seidler, H., & Corona-Strauss, F. I. (2013). Neural correlates of listening effort related factors: Influence of age and hearing impairment. *Brain Research Bulletin, 91*, 21-30.

Best, V., Keidser, G., Buchholz, J. M., & Freeston, K. (2015). An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment. *International Journal of Audiology, 54*(10), 682-690.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58(3), 306–324.

Friesen, L. M., Shannon, R. V., Baskent, D., and Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America, 110*(2), 1150–1163.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 252-269.

Gosselin, P. A., & Gagné, J.P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*(3), 944-958.

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology, 33*, 457–461.

Hawkins, D. B., & Yacullo, W. S. (1984). Signal-to-noise ratio advantage of binaural hearing aids and directional microphones under different levels of reverberation. *Journal of Speech and Hearing Disorders, 49*(3), 278-286.

Herrmann, C. S., Fründ, I., & Lenz, D. (2010). Human gamma-band activity: a review on cognitive and behavioral correlates and network models. *Neuroscience & Biobehavioral Reviews, 34*(7), 981-992.

Hornsby, B. W., Werfel, K., Camarata, S., & Bess, F. H. (2014). Subjective fatigue in children with hearing loss: Some preliminary findings. *American Journal of Audiology, 23*(1), 129-134.

Hua, H., Emilsson, M., Ellis, R., Widén, S., Möller, C., & Lyxell, B. (2014). Cognitive skills and the effect of noise on perceived effort in employees with aided hearing impairment and normal hearing. *Noise and Health, 16*(69), 79-88.

Hua, H., Anderzén-Carlsson, A., Widén, S., Möller, C., & Lyxell, B. (2015). Conceptions of working life among employees with mild-moderate aided hearing impairment: A phenomenographic study. *International Journal of Audiology, 54*(11), 873-880.

Karrasch, M., Laine, M., Rapinoja, P., & Krause, C. M. (2004). Effects of normal aging on event-related desynchronization/synchronization during a memory task in humans. *Neuroscience letters, 366*(1), 18–23.

Keidser, G., Dillon, H., Mejia, J., & Nguyen, C.-V. (2013). An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function. *International Journal of Audiology, 52*(11), 795-800.

Klimesch, W. (1996). Memory processes, brain oscillations and EEG synchronization. *International Journal of Psychophysiology, 24*(1), 61-100.

Klimesch, W. (1999). EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Research Reviews, 29*(2), 169-195.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences, 16*(12), 606-617.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291-300.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology, 2012,* 1-11.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2014). The influence of informational masking on speech perception and pupil response in adults with hearing impairment. *The Journal of the Acoustical Society of America, 135*(3), 1596-1606.

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *International Journal of Audiology, 36*(3), 155-164.

Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *International Journal of Audiology, 45*(9), 503-512.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*(1), 23-34.

Kuper, H., Singh-Manoux, A., Siegrist, J., & Marmot, M. (2002). When reciprocity fails: effort–reward imbalance in relation to coronary heart disease and health functioning

within the Whitehall II study. *Occupational and Environmental Medicine, 59*(11), 777-784.

Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology, 44*(3), 131-143.

Leiberg, S., Lutzenberger, W., & Kaiser, J. (2006). Effects of memory load on cortical oscillatory activity during auditory pattern working memory. *Brain research, 1120*(1), 131–140.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *International Journal of Audiology, 42*, Supp1, 49-58.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology, 53*(7), 433-440.

McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., . . . Graham, P. L. (2016). Monitoring alpha oscillations and pupil dilation across the performance-intensity function. *Frontiers in Psychology,* 7: 745.

Mehta, R. K., & Agnew, M. J. (2012). Influence of mental workload on muscle endurance, fatigue, and recovery during intermittent static work. *European Journal of Applied Physiology, 112*(8), 2891-2902.

Mele, M. L., & Federici, S. (2012). Gaze and eye-tracking solutions for psychological research. *Cognitive Processing, 13*(1), 261-265.

Nachtegaal, J., Kuik, D. J., Anema, J. R., Goverts, S. T., Festen, J. M., & Kramer, S. E. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. *International Journal of Audiology, 48*(10), 684-691.

Obleser, J., & Weisz, N. (2012). Suppressed alpha oscillations predict intelligibility of speech and its acoustic details. *Cerebral Cortex, 22*(11), 2466–2477.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *The Journal of Neuroscience, 32*(36), 12376-12383.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2010). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*.

Pesonen, M., Björnberg, C. H., Hämäläinen, H., & Krause, C. M. (2006). Brain oscillatory 1 - 30 Hz EEG ERD/ERS responses during the different stages of an auditory memory search task. *Neuroscience Letters, 399*(1), 45–50.

Peters, M. L., Godaert, G. L., Ballieux, R. E., van Vliet, M., Willemsen, J. J., Sweep, F. C., & Heijnen, C. J. (1998). Cardiovascular and endocrine responses to experimental stress: effects of mental effort and controllability. *Psychoneuroendocrinology, 23*(1), 1-17.

Petersen, E. B., Wöstmann, M., Obleser, J., Stenfelt, S., & Lunner, T. (2015). Hearing loss impacts neural alpha oscillations under adverse listening conditions. *Frontiers in Psychology, 6*: 177.

Pichora-Fuller, K. (2003). Cognitive aging and auditory information processing. *International Journal of Audiology, 42,* Supp 2, 26-32.

Pichora-Fuller, M. K. (2006). Perceptual effort and apparent cognitive decline: Implications for audiological rehabilitation. *Seminars in Hearing, 27,* 284 –293.

Pichora-Fuller, K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: implications for hearing aid fitting and audiologic rehabilitation. *Trends in Amplification, 10*(1), 29-59.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., ... & Naylor, G. (2016). Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing, 37,* 5S-27S.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2014). R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-117. *Available at h ttp://CRAN. R-project. org/package= nlme*.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560-569.

Power, J. D., & Petersen, S. E. (2013). Control-related systems in the human brain. *Current Opinion in Neurobiology, 23,* 223–228.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology, 20*(3), 241-248.

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology, 47,* Supp2, 99-105.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology, 23*(8), 577-589.

Scharinger, C., Kammerer, Y., & Gerjets, P. (2015). Pupil Dilation and EEG Alpha Frequency Band Power Reveal Load on Executive Functions for Link-Selection Processes during Text Reading. *PLoS ONE, 10*(6), e0130608.

Schneider, B. A., Pichora-Fuller, K., & Daneman, M. (2010). Effects of senescent changes in audition and cognition on spoken language comprehension. In S. Gordon-Salant, D. R. Frisina, A. N. Popper & R. R. Fay (Eds.), Springer handbook of auditory research: The aging auditory system (vol. 34,). New York: Springer.

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology, 45*(5), 679-687.

Siegrist, J. (1996). Adverse health effects of high-effort/low-reward conditions. *Journal of Occupational Health Psychology, 1*(1), 27-41.

Smits, C., & Festen, J. M. (2013). The interpretation of speech reception threshold data in normal-hearing and hearing-impaired listeners: II. Fluctuating noise. *The Journal of the Acoustical Society of America, 133*(5), 3004-3015.

Souza, P., & Arehart, K. (2015). Robust relationship between reading span and speech recognition in noise. *International Journal of Audiology, 54*(10), 705-713.

Strauß, A., Wöstmann, M., & Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience,* 8: 350.

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging, 24*(3), 761.

Ward, L. M. (2003). Synchronous neural oscillations and cognitive processes. *Trends in Cognitive Sciences, 7*(12), 553-559.

Weinstein, B. E., & Ventry, I. M. (1982). Hearing Impairment and Social Isolation in the Elderly. *Journal of Speech, Language, and Hearing Research, 25*(4), 593-599.

Weisz, N., Hartmann, T., Mu¨ller, N., & Obleser, J. (2011). Alpha rhythms in audition: cognitive and clinical perspectives. *Frontiers in Psychology, 2*, 73.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology,* 7: 345.

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing loss in older adulthood what it is and how it interacts with cognitive performance. *Current Directions in Psychological Science, 14*(3), 144-148.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing, 36*(4), 153-165.

Wolak, M. E., Fairbairn, D. J., & Paulsen, Y. R. (2012). Guidelines for estimating repeatability. *Methods in Ecology and Evolution, 3*(1), 129-137.

Wöstmann, M., Herrmann, B., Wilsch, A., & Obleser, J. (2015). Neural alpha dynamics in younger and older listeners reflect acoustic challenges and predictive benefits. *The Journal of Neuroscience, 35*(4), 1458-1467.

Wouters, J., & Berghe, J. V. (2001). Speech recognition in noise for cochlear implantees with a two-microphone monaural adaptive noise reduction system. *Ear and Hearing, 22*(5), 420-430.

Zekveld, A. A., Heslenfeld, D. J., Festen, J. M., & Schoonhoven, R. (2006). Top–down and bottom–up processes in speech comprehension. *Neuroimage, 32*(4), 1826-1836.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing, 31*(4), 480-490.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing, 32*(4), 498-510.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: neural correlates of pupil size as a measure of cognitive listening load. *Neuroimage, 101*, 76-86.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*(3), 277-284.

Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., & Rönnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Frontiers in Neuroscience*, 8: 88.

Chapter 5 Physiological measures of listening effort and interactions with working memory capacity: a short communication

Kelly Miles[1,2,3], Isabelle Boisvert[1,2], Catherine McMahon[1,2], and Björn Lyxell[3]

[1] Centre for Implementation of Hearing Research, Macquarie University, Sydney, Australia

[2] The HEARing Cooperative Research Centre, Melbourne, Australia

[3] Linnaeus Centre for HEaring And Deafness (HEAD), Swedish Institute for Disability Research, Linköping University, Sweden

5.1 Introduction

This short communication investigated under which conditions working memory capacity (reading span task) interacted with the physiological measures of listening effort. The terms 'offline' and 'online' measures of working memory capacity are often used in the literature to reference *when* a measure of working memory capacity was assessed. An offline measure of working memory capacity refers to an assessment of working memory capacity that was performed independently (time separated) of physiological testing. The working memory capacity result is subsequently used as a variable to assess its impact on physiological data. On the other hand, an online measure of working memory capacity indicates that the assessment was concurrently measured (e.g., EEG synchronisation *during* a working memory task; Stam, van Walsum & Micheloyannis, 2002). While this short communication does not advocate for one method over the other, it is necessary to make this distinction in order to interpret the varying results in the literature.

In the pupillometry literature, varying results regarding interactions with working memory capacity have been reported. For example, Koelewijn, Zekveld, Festen, Rönnberg, and Kramer (2012) found that greater working memory capacity measured in an offline task was associated with increased pupil size. In a similar paradigm, Zekveld and Kramer (2014) found no correlation between working memory capacity and pupil size. Measuring pupil size online during a working memory task also shows disparate findings. Heitz and colleagues (2008) demonstrated that greater working memory capacity resulted in smaller pupil diameters, whereas Wendt et al. (2016) found that a working memory task was positively correlated with pupil size during a sentence recognition task (i.e., greater working memory capacity was associated with larger pupil diameters).

Alpha oscillations and working memory capacity have generally been examined during online working memory tasks in the visual domain. For example, studies have shown that

alpha power increases with increased working memory load during a Sternberg paradigm (Jensen, Gelfand, Kounios, & Lisman, 2002; Tuladhar et al., 2007). Adapting the Sternberg task for the auditory domain, Obleser et al. (2012) demonstrated that both spectral resolution and increased working memory load resulted in additive alpha power. That is, when spectral resolution and working memory load was greatest, alpha power change was also greatest.

Chapter 3 in the current thesis found that on average, individuals with higher working memory capacity rated listening to 6-channel vocoded sentences as more effortful compared to 16-channel vocoded sentences. This was not the case for individuals with lower working memory capacity. As the same participants in Chapter 3 also participated in a physiological testing session to assess pupil dilation and alpha power change (Chapter 4), we aim to assess whether an offline measure of working memory capacity predicted pupil size and alpha power change during the sentence recognition task.

Based on the previous literature, we hypothesise that greater working memory capacity will be associated with greater alpha power, however as various studies have found disparate results relating to working memory capacity and pupil size, whether a relationship will exist is an open question.

5.2 Methods

Comprehensive details of the Methods used can be found in Miles et al. (2017), Chapter 4 in the current thesis.

Participants

Participants (12 women, 7 men) had a mean age of 27 years (SD = 4.28, range = 22-34 years) and normal hearing.

Speech materials and background noise

Bamford-Kowal-Bench sentences adapted for Australian-English (Bench & Doyle, 1979) were recorded by a female speaker. The sentences and background noise (4-talker babble) were vocoded using custom MATLAB scripts.

Physiological measures

The pupil and EEG recordings were simultaneously recorded at 1000 Hz.

Working memory capacity

The reading span test was used as a measure of working memory capacity (Daneman & Carpenter, 1980). Short sentences were presented visually on a computer screen, in three meaningful segments separated by 50 ms (e.g., "the dad", "hugged", "the daughter"), in blocks of three, four, and five sentences. Each segment appeared for 800 ms. After each sentence, the participant had 1.75 seconds to determine whether the sentence made sense or not by pressing 'Y' or 'N' on a keyboard. After each block of sentences, participants were asked to repeat either the first or the last words of each sentence. The order of presentation was randomised. Scoring was based on the total number of words correctly recalled (Rönnberg, Arlinger, Lyxell, & Kinnefors, 1989). For further analysis, a median split (67%) was also conducted on the data to form two groups consisting of individuals with higher working memory capacity and lower working memory capacity, per the analysis conducted in Chapter 3.

5.3 Statistical analyses

All analyses were performed in R version 3.2.1 using the nlme package (Pinheiro, Bates, DebRoy, & Sarkar, 2014). Linear mixed-effects (LME) models with a random intercept for individual were used for all analyses to control for repeated measures over different levels of the factors on individuals. Only interaction terms were assessed. A P value <0.05 was considered significant. Independent variables were actual performance and channel-vocoding, and the dependent variables were pupil dilation and alpha power. Two models were built for

each measure, one with working memory capacity as a continuous predictor variable, and the other with a high/low working memory capacity as a dichotomous covariate. Only sentences that were partially or correctly answered were included in the analysis, as this ensured participants were on task.

5.4 Results

The summary results can be found in Table 1. The results revealed no significant main effects or interactions for any predictor variables across models, as shown in Table 2. Although not statistically significant, the performance x high/low working memory interaction showed on average, participants in the lower working memory capacity group had larger pupil diameters than the higher working memory capacity group, but smaller pupil diameters for channel-vocoding. Conversely, alpha power was lower in the performance x high/low working memory interaction for the lower working memory capacity group, and higher for channel-vocoding, compared to the higher working memory capacity group.

Table 1. Summary statistics of the measures.

	High working memory capacity		**Low working memory capacity**	
Measure	Mean	SD	Mean	SD
Actual performance	60.37	17.9	57.76	17.54
Working memory capacity	73.79	5.31	54.03	9.78
Maximum pupil size	14.31	15.4	10.85	11.58
Mean alpha power	133.84	408.61	127.15	378.11

Table 2. Results for the Linear Mixed-Effects Models.

	Pupil size			**Alpha power**		
Model	Estimate	95% CI	*p* value	Estimate	95% CI	*p* value
Actual performance x WMC	-0.001	[-0.005, 0.003]	0.127	0.080	[-0.126, 2.862]	0.690
Vocoding x WMC	0.072	[-0.289, 0.434]	0.228	0.508	[-0.687, 1.702]	0.553
Actual performance x Vocoding x WMC	-0.002	[-0.007, 0.004]	0.509	-0.009	[-0.027, 8.963]	0.323
Actual performance x HLWMC	-0.014	[-0.17, 0.142]	0.321	-0.068	[-5.093, 4.958]	0.192
Vocoding x HLWMC	-0.015	[-0.838, 0.807]	0.142	0.670	[-25.86, 27.199]	0.115
Actual performance x Vocoding x HLWMC	0.003	[-0.011, 0.017]	0.689	0.079	[-0.364, 0.522]	0.727

5.5 Discussion

This short communication aimed to assess whether working memory capacity, as a continuous or dichotomous covariate variable, predicted pupil dilation or alpha power changes across levels of vocoding and performance. The results indicated that neither working memory capacity variable interacted with the physiological measures.

The pupillometric findings are in agreement with Zekveld and Kramer (2014) who found no association between working memory capacity and pupil size change during a sentence recognition task. However this outcome is contrary to Koelewijn, Zekveld, Festen, Rönnberg, et al. (2012) who reported a significant association between an offline measure of working memory capacity and pupil size change during a speech recognition task. This observed difference may however be attributed to the different baseline corrections applied to the data, whereby their study used absolute change from baseline whereas the current study used relative percent change from baseline. As demonstrated in Chapter 6, type of baseline correction can significantly alter the results of the same pupillometric dataset.

While many studies have shown that online measures of working memory capacity modulate alpha power, this was not the case for the offline measure of working memory capacity used in the current study. It is possible that this may be due to the different working memory assessments used between the studies, as a Sternberg paradigm (most frequently used in online working memory tasks) is a relatively simple test compared to a reading span task. That is, recalling whether a probe digit was in the preceding 2-6 digits is different from recalling a semantically correct or incorrect sentence and remembering the first or final nouns (as was the case in the reading span used in the current study). It may also be that offline measures of working memory capacity are not predictive of alpha power change during a sentence recognition task. While pupillometric studies have demonstrated associations with online reading/ listening span tasks (Heitz et al., 2008; Wendt et al., 2016), to the authors'

knowledge, alpha power change has not been examined during an online reading/ listening span assessment in different types of background noise. This is an important area of future research.

5.6 References

Bench, R., & Doyle, J. (1979). *The Bamford-Kowal-Bench/Australian version (BKB/A) Standard Sentence Lists.* Carlton, Victoria: Lincoln Institute.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466.

Heitz, R. P., Schrock, J. C., Payne, T. W., & Engle, R. W. (2008). Effects of incentive on working memory capacity: Behavioral and pupillometric data. *Psychophysiology, 45*(1), 119-129.

Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral Cortex, 12*(8), 877-882.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology, 2012*, 1-11.

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., & Lyxell, B. (2017). Objective assessment of listening effort: co-registration of pupillometry and EEG. *Trends in Hearing.*

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *The Journal of Neuroscience, 32*(36), 12376-12383.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2014). R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-117. Available at http://CRAN. R-project. org/package= nlme.

Rönnberg, J., Arlinger, S., Lyxell, B., & Kinnefors, C. (1989). Visual Evoked PotentialsRelation to Adult Speechreading and Cognitive Function. *Journal of Speech, Language, and Hearing Research, 32*(4), 725-735.

Stam, C. J., van Walsum, A. M. V. C., & Micheloyannis, S. (2002). Variability of EEG synchronization during a working memory task in healthy subjects. *International Journal of Psychophysiology, 46*(1), 53-66.

Tuladhar, A. M., Huurne, N. t., Schoffelen, J. M., Maris, E., Oostenveld, R., & Jensen, O. (2007). Parieto-occipital sources account for the increase in alpha activity with working memory load. *Human Brain Mapping, 28*(8), 785-792.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, 7: 345.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*(3), 277-284.

Chapter 6 Pupillometry as a measure of listening effort: methodological considerations for data and statistical analysis

Kelly Miles[1,2,3], Isabelle Boisvert[1,2], Catherine M. McMahon[1,2], Nay San[4,5], Kenneth Beath[6], Timothy Beechey[1,7], and Björn Lyxell[3]

[1] Centre for Implementation of Hearing Research, Macquarie University, Sydney, Australia

[2] The HEARing Cooperative Research Centre, Melbourne, Australia

[3] Linnaeus Centre for HEaring And Deafness (HEAD), Swedish Institute for Disability Research, Linköping University, Sweden

[4] Australian National University, Canberra, Australia

[5] The ARC Centre of Excellence for the Dynamics of Language, Canberra, Australia

[6] Department of Statistics, Australian Hearing Hub, Macquarie University, Sydney, Australia

[7] National Acoustic Laboratories, Sydney, Australia

6.1 Introduction

Since the early work of Hess and Polt (1964), pupillometric measures have been used across multiple disciplines as a measure of exerted effort. In hearing research, pupillometry is now one of the most common tools to assess the amount of effort exerted when listening with a hearing impairment and/ or in adverse environments. In a seminal study based largely on early literature exploring pupil dilation in response to mental activity such as arithmetic and memory processes, Kramer and colleagues (1997) investigated whether pupil dilation in response to changes in listening difficulty varied as a function of hearing loss. Following this, over 25 studies have examined how different variables such as age, hearing loss, background noise, performance levels, spectral quality, spatial cues and/ or linguistic manipulations affect the pupil response. In general, it has typically been found that as listening demands become more challenging, the pupil diameter increases. This has been interpreted as reflecting increased listening effort.

Past studies examining the pupil response to listening challenges have informed the field, but there is currently a lack of guidance regarding best practice for data analysis and statistical modelling. Multiple methods have been used across studies to scale and baseline-correct data, along with disparate statistical approaches, making it challenging to compare results between studies. Importantly, different processing and statistical approaches may yield different results within the same dataset, thus affecting our understanding of how different variables, such as population demographics and task manipulations, may affect listening effort.

Varying data scaling options

Scaling data is used to control for differences in pupil diameter across individuals, or to remove systematic within-trial variation. Traditionally, and most frequently in the pupillometry/ listening effort (P-LE) literature, data entered into a statistical model is not

scaled. Those studies that have scaled data generally used dispersion as the scaling factor (e.g., range scaling which is also known as feature scaling; Wendt, Dau, & Hjortkjær, 2016) or an average value (e.g., within-trial mean scaling; Kuchinsky et al., 2013). A comprehensive review and evaluation of centering, scaling and transforming data can be found in van den Berg, Hoefsloot, Westerhuis, Smilde, and van der Werf (2006). However, for the purposes of this paper, a brief overview of the scaling options commonly applied in the P-LE literature is provided here.

Range scaling [$(x_i - \min)/ \max - \min$] uses the range within a trial as the scaling factor to control for differences in pupil diameter across individuals and trials. As range scaling results in a value between 0 and 1, it is particularly useful for multivariate analysis comparing measures with different units. It is, however, sensitive to inflation errors and outliers as the range is calculated from two data points (the maximum and minimum value) within a trial - which may themselves be outliers (van den Berg et al., 2006).

Mean scaling [$x_i / \bar{x}$] differs from range scaling as it uses an average value as the scaling factor instead of a measure of dispersion reflecting relative change. This method has been used by Kuchinsky et al. (2014; 2013) to control for significant correlations between trial averages and standard deviations.

Varying baseline correction options

Baseline correction is carried out to improve signal margin. Best practice includes a baseline period that is minimally contaminated by preparatory responses and/or the preceding trial (Luck, 2014). Across studies, baseline periods may differ in time and stimulus (e.g., a 500 ms baseline in noise versus a 1000 ms baseline in quiet) and the type of baseline correction.

The vast majority of studies use an absolute change from baseline correction [x-BL] where change is identified in terms of subtracting the baseline value from a region of interest. Only

few studies have used relative percent change from baseline [(x-BL)/BL*100] in the P-LE literature (Miles et al., 2017; Wagner, Pals, de Blecourt, Sarampalis, & Başkent, 2016; Wagner, Toffanin, & Başkent, 2016) although its use has been recommended for statistical analyses of pupillometric data as it removes additional inter-individual variability (Lemercier et al., 2014). It is however noted that large data points can skew distributions when using a relative percent change, which may result in lower statistical power (Vickers, 2001).

Varying statistical modelling options

The range of statistical models and software packages used to analyse physiological data has rapidly evolved over the years leading to a variety of statistical options when analysing data.

Repeated-measures ANOVA model (rmANOVA) - One of the most common statistical approaches in the P-LE literature is the rmANOVA which tests for differences in mean scores and distributions across factors. Majority of P-LE studies have employed a within-subjects design where each participant is exposed to all experimental conditions. Including a repeated measure controls for differences between participants and factors which reduces model variance. The rmANOVA partitions the variance allowing for correlation within subjects and then performs the analysis using ANOVA.

Linear mixed-effects model (LME) - LMEs are an extension of linear regression where both fixed and random effects are modelled in a single equation. Fixed effects are considered to be the same for all subjects, as in linear regression, whereas random effects vary across subjects, usually assumed to have a normal distribution. Consequently, they allow for the lack of independence between observations within a subject and allow more complex data than the rmANOVA.

Bayesian model - In recent years, Bayesian statistical modelling has become increasingly popular in many fields, including in the analysis of pupillometric data (Allen et al., 2014;

Cavanagh, Wiecki, Kochar, & Frank, 2014). In the past, the programming skill and computational expense required to implement Bayesian models made its application largely prohibitive. However, the introduction of modern techniques including (i) Markov chain Monte Carlo (MCMC) algorithms, particularly implementations of Gibbs sampling (Plummer, 2003) and Hamiltonian Monte Carlo (Carpenter et al., 2016); and (ii) fast Bayesian approximation algorithms (Rue et al., 2014) have greatly simplified Bayesian analyses making it accessible to the wider-research community.

The previously discussed statistical models use point estimates and confidence intervals to draw inferences. Bayesian analysis diverges from these models in that it uses conditional probability to obtain probability distributions for model parameters. Because Bayesian methods treat model parameters (e.g., slopes and intercepts) as random variables (in contrast to frequentist methods which instead treat data as a random variable), all model parameters and estimates based on parameters (e.g., mean estimates) have full posterior distributions rather than single point estimates. In this way Bayesian models can more explicitly communicate the uncertainty inherent in estimation and prediction. Bayesian methods are also particularly suitable for small sample sizes and non-normally distributed data because inference is not based upon a sampling distribution which is assumed to be asymptotically accurate only with very large sample sizes.

6.2 Aim

To understand how these varying data processing and statistical approaches can influence the interpretation of results, this study compares the effect of using different data scaling methods, baseline correction options, and statistical models in a single pupillometric dataset (from Miles et al., 2017, Chapter 4). These comparisons aim to provide empirical evidence regarding the replicability and reliability of published studies, and the potential limitations of comparing across published datasets. There is an urgent need to address these factors if

pupillometry is to be considered a viable research tool to assess listening effort, and importantly, how these issues influence the candidacy of pupillometry as a clinical tool.

6.3 Materials and methods

Participants

Data from 23 participants recruited as part of a larger study (Miles et al., 2017; Chapter 4) are included in the current analysis. Chapter 4 only included participants who had greater than 65% of trials accepted for both pupil dilation *and* EEG measures. The current study is only concerned with pupil dilation, and therefore more participants had greater than 65% of accepted trials for the pupil dilation measure. Participants (13 female, 10 men) had a mean age of 27.47 years (SD = 4.01, range = 20-34 years). Participants had present distortion-product otoacoustic emissions between 1-4 kHz. The Macquarie University Human Research Ethics Committee approved the study.

Study design and protocol

The study design and data used in this paper are comprehensively described in Miles et al. (2017). In brief, pupil size was measured during a sentence recognition task, where listening difficulty was manipulated by parametrically varying spectral resolution (16- and 6-channel channel-vocoding) and signal-to-noise ratios (SNRs; individually-measured speech reception thresholds (SRT) for 50% and 80% performance accuracy). The four conditions each contained 55 sentences and were randomly presented. Table 1 shows the mean and standard deviations of the SNRs for each condition. Results from Miles et al., (2017) demonstrated that there were inconsistencies between the target SRTs (obtained in a prior session) and the true performance obtained during the pupil recording session. For example, where the target SRT was 80% performance accuracy, participants scored an average of 64.7% (36 – 93%). As such, two models were developed to assess the predictor variables: 1) channel-vocoding and target SRT, and 2) channel-vocoding and true performance. The pupil response was

sensitive to changes in channel-vocoding (pupil diameter was larger for the 6-channel condition) and true performance (pupil diameter was smaller as performance approached 100%), but not to the target SRT. For an in-depth discussion of model parameters, see Miles et al. (2017). The models in the current paper are for illustrative purposes, and therefore, to simplify model terms, only channel-vocoding and SRT are used as factors. Similarly, while different studies have used different pupil parameters such as mean or maximum size, and latency to maximum size, the current paper will only include mean pupil size. The analyses are therefore different to those included in Chapter 4 which used maximum pupil size to further demonstrate the analysis types across different pupillometry metrics. Note that all data, including mean, maximum and latency to maximum pupil size is available in the repository.

Table 1. Means and standard deviations of SNRs obtained during the speech recognition task.

SRT %	*50*		*80*	
Vocoding	*6*	*16*	*6*	*16*
Mean (dB)	0.9	-1.8	4.0	0.5
SD (dB)	1.7	1.5	1.9	1.6

Data preparation

Offline, single-trials were processed with Dataviewer software (version 1.11.1), and compiled into single-trial pupil-diameter waveforms (-1 s to 5 s). Custom MATLAB scripts were developed for blink identification, interpolation, and data smoothing. This code has been made available on a public repository (https://www.github.com/mileskm/analyses). Blinks were identified as pupil sample sizes smaller than three standard deviations below the mean pupil diameter, on a trial-by-trial basis. Trials with more than 15% of the samples detected in blink were rejected. Linear interpolation was used to reduce lost samples and artefacts from

blinks for the remaining accepted trials (Siegle, Ichikawa, & Steinhauer, 2008; Zekveld, Kramer, & Festen, 2010). Samples detected in blink were interpolated from 66 samples preceding the onset to 132 samples following the offset of a blink. Note that this interpolation covers a much larger range of samples due to the higher sampling rate of the EyeLink eyetracker used in this study (1000 Hz) compared to many studies that use a 60 Hz sampling frequency (e.g., Kuchinsky et al., 2013; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014). After interpolation, data were smoothed using a 5-points moving average. Means were calculated over the regions of interest including the baseline in noise (0-1 s) and the encoding period (2-6 s).

The naming conventions of the pupil measures available in the repository are outlined in Table 2. In addition to the pupil data, basic demographics and behavioural test results have also been made available. Datawrangling was performed using the R-tidyverse package (Wickham, 2017). In the repository, a different file is used for each statistical model due to the necessary differences in data aggregation (Table 3). Each file contains the participant identifier (SubjID), SRT (50 or 80%), and channel-vocoding (16- and 6-channel), along with the pupil measurements.

Table 2. Data processing naming codes. Note that while 'raw data' has been used here, the data has been pre-processed (see Data Preparation section, above). Raw here refers to the data not having undergone scaling.

Naming code	**Processing steps**
raw_abs	- Data has not undergone any transformation - Baseline corrected using absolute change from baseline [x-BL]
rangeScale_abs	- Data has undergone range scaling [(x – min)/ max - min] - Baseline corrected using absolute change from baseline [x-BL]
meanScale_abs	- Data has undergone mean scaling [xi / $\bar{x}$] - Baseline corrected using absolute change from baseline [x-BL]
raw_perc	- Data has not undergone any transformation

	- Baseline corrected using relative percent change from baseline [(x-BL)/BL*100]

Table 3. Aggregated model naming codes

Model	**Aggregation**
rmANOVA	Averaged over participant and the factors SRT and Vocoding
LME & Bayes	Averaged over participant, trial, and the factors SRT and Vocoding

6.4 Statistical methods

All analyses were performed in R version 3.3.3 (R Core Team, 2008). Only main effects models were developed. rmANOVA models were analysed using the R-afex package (Singmann, Bolker, & Westfall, 2015) and R-lsmeans package (Lenth & Hervé, 2013). The error term was set as subject error by the within subject error (SRT and Vocoding). LME models were analysed using the R-nlme package (Pinheiro, Bates, DebRoy, & Sarkar, 2014). Random intercepts for individuals were used to control for repeated measures over different levels of the factors on individuals. Bayes models were analysed using the R-INLA package (Martins, Simpson, Lindgren, & Rue, 2013; Rue et al., 2014). Bayes models were set to use a gaussian distribution and default regularising priors.

Confidence intervals (CIs) are reported instead of p-values for the rmANOVA and LME model output. For the Bayesian models, credible intervals are reported. Reporting CIs facilitates comparison between statistical models (p-values are not typically calculated for Bayesian statistics), and additionally provide information about the direction and strength of the effect (Shakespeare, Gebski, Veness, & Simes, 2001). To aid interpretability, statistically significant results have been asterisked in the summary table. Effect sizes were calculated using Hedges' g.

6.5 Results

The different data processing methods considerably affected the input data for statistical analysis, as shown in Figure 1 (rmANOVA data) and Figure 2 (LME/ Bayes data). Summary descriptive statistics are presented in Table 4, grouped by statistical model and data processing approach. Note that aggregating data for the different statistical models results in a different number of data points between rmANOVA (23 participants x 4 conditions = 92) and the LME/ Bayes models (23 participants x 4 conditions x +/-55 sentences = 4578). As illustrated in Figure 3, there was a large amount of inter-individual variability.

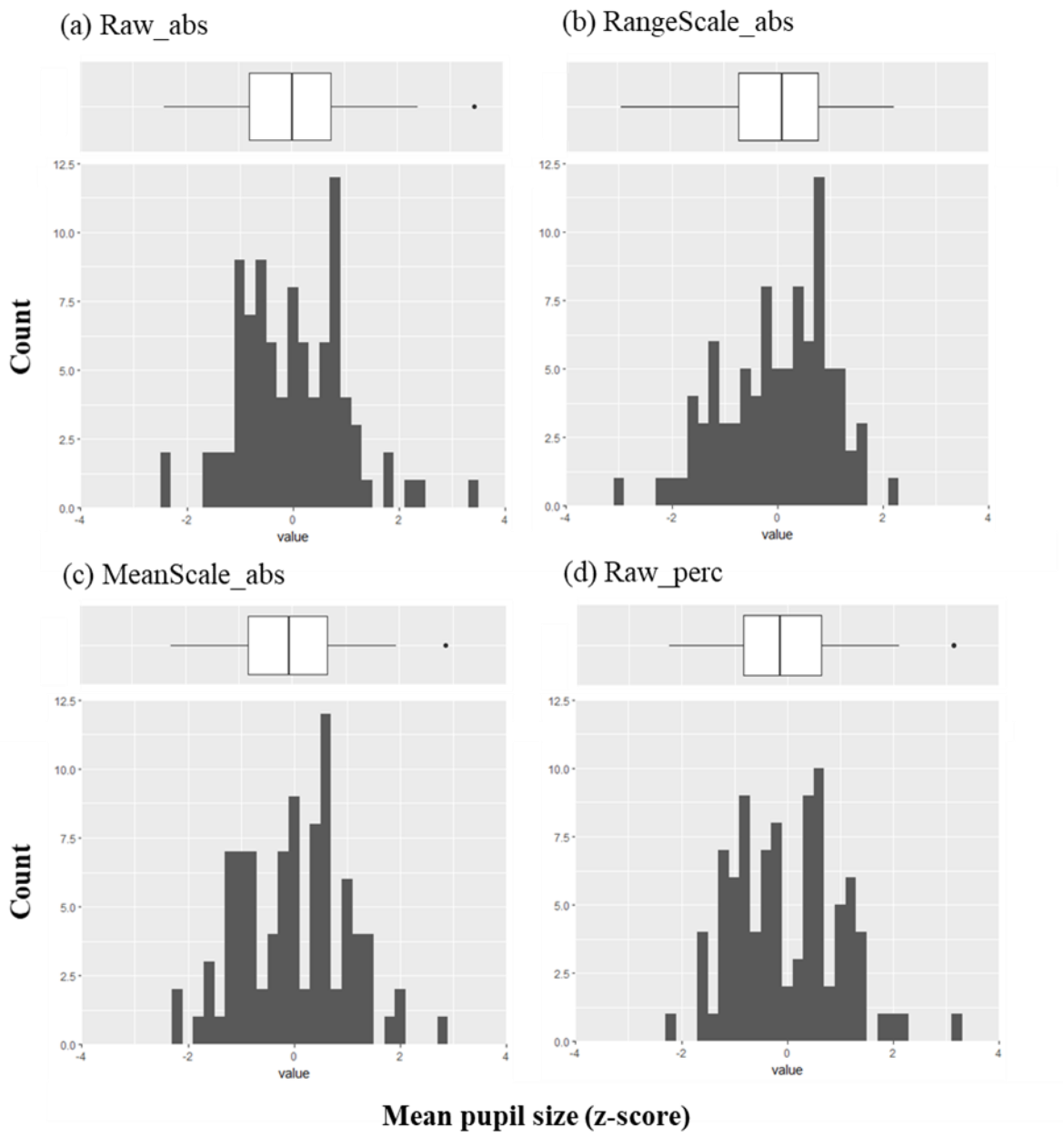

Figure 1. Histograms (binwidth .2) and boxplots of a single pupil size dataset aggregated for rmANOVA and processed as follows: a) non-scaled and absolute baseline correction, b) range-scaled and absolute baseline correction, c) mean-scaled and absolute baseline correction, and d) non-scaled and relative baseline correction. Z-scores are used to facilitate comparison on a common scale.

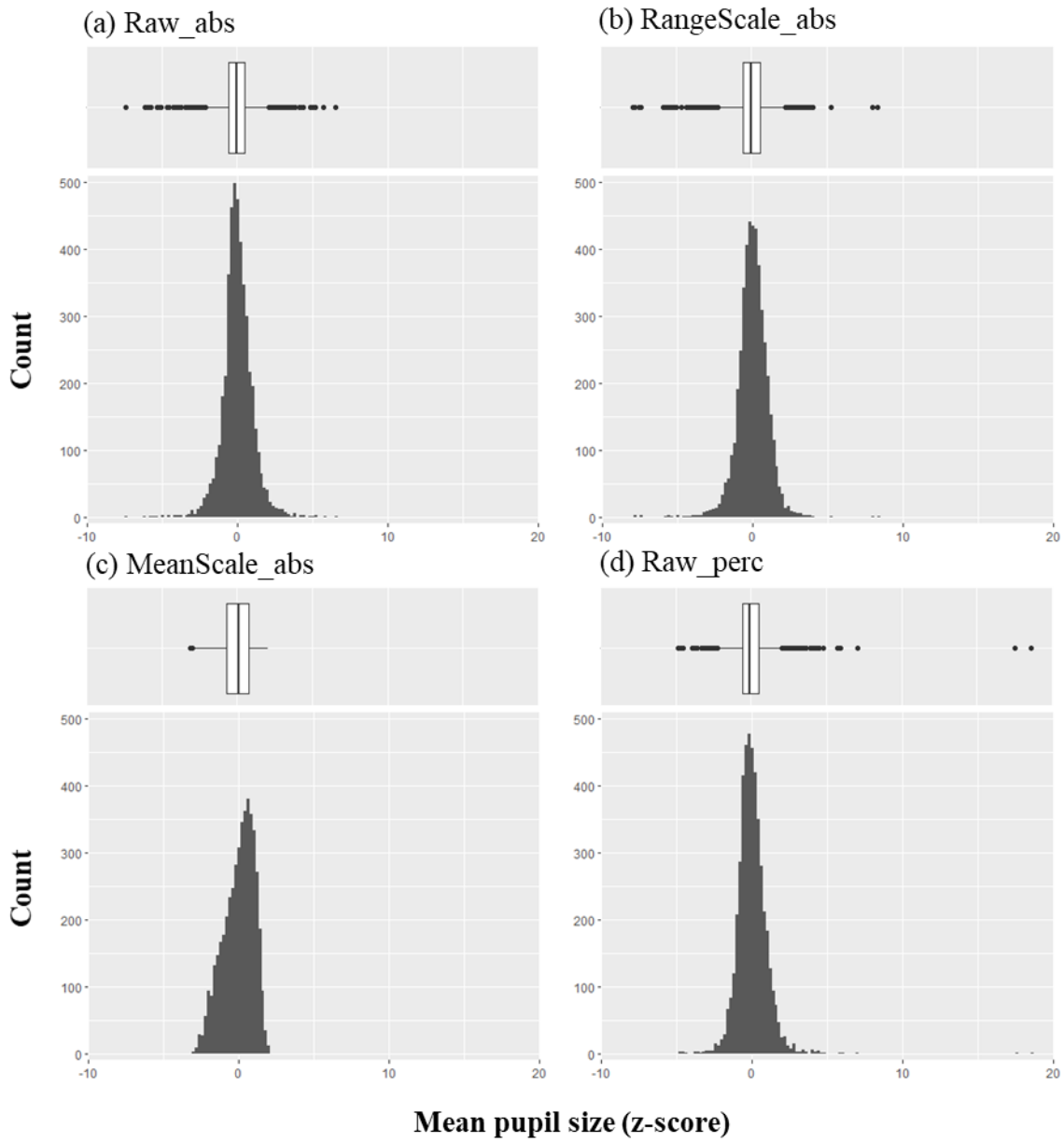

Figure 2. Histograms (binwidth .2) and boxplots a single pupil size dataset aggregated for LME/ Bayes analyses and processed as follows: a) non-scaled and absolute baseline correction, b) range-scaled and absolute baseline correction, c) mean-scaled and absolute baseline correction, and d) non-scaled and relative baseline correction. Z-scores are used to facilitate comparison on a common scale.

Data scaling

As expected, the most notable difference between the raw and scaled data was the shape of the distributions (compare Figure 2(a) and Figure 2(b) as an example). Data aggregation for rmANOVA compared to LME/ Bayes models also affected the distributions (compare Figure 1 with Figure 2). Range scaled data appeared relatively robust to data aggregation, although both distributions are moderately negatively skewed (Table 4; Skew). As illustrated by the box-plots in Figure 2, the LME data aggregation across data scaling methods (aside from range-scaled data) show a considerable number of outliers in the distribution tails. The raw and scaled data show relatively comparable means across the aggregated data while the measures of dispersion and range show large differences.

Baseline correction

The different baseline corrections were compared (absolute change and percent change) on the raw data only (raw_abs and raw_perc). In the rmANOVA aggregation, the distributions were comparable, with a moderate positive skew. Absolute baseline correction led to a more platykurtic distribution than percent change (see Figure 1(a) and Figure 2(d)). The absolute change from baseline correction had almost perfect symmetry with the LME/ Bayes aggregation, although the distribution was more platykurtic compared to the rmANOVA aggregation (Figure 1(a) and Figure 2(a)). Percent baseline correction substantially transformed the shape of the distribution between data aggregations, resulting in both a highly skewed and platykurtic distribution for the LME/ Bayes aggregation (Figure 2(d)). Extreme data points were emphasised in the LME/ Bayes aggregation for percent change from baseline correction, resulting in two data points greater than 200%.

Table 4. Summary statistics of the different statistical models, scaling methods, and baseline corrections collapsed across channel-vocoding and SRT, applied to a single pupillometric dataset.

	Compared approach	n	Mean	SD	Median	SE	Min	Max	Range	Skew	Kurtosis
rmANOVA	raw_abs	92	109.67	93.12	111.65	9.71	-114.50	431.40	545.90	0.41	0.70
	rangeScale_abs	92	0.14	0.09	0.15	0.01	-0.13	0.35	0.48	-0.40	-0.37
	meanScale_abs	92	0.05	0.04	0.04	0.00	-0.04	0.15	0.19	0.06	-0.32
	raw_perc	92	5.73	3.95	5.18	0.41	-3.12	18.15	21.26	0.30	-0.18
LME/ Bayes	raw_abs	4578	107.17	252.71	96.20	3.73	-1764.62	1754.68	3519.30	-0.01	4.33
	rangeScale_abs	4578	0.14	0.28	0.19	0.00	-0.73	0.73	1.46	-0.50	-0.47
	meanScale_abs	4578	0.05	0.11	0.05	0.00	-0.79	0.94	1.73	-0.48	7.14
	raw_perc	4578	5.65	12.21	4.81	0.18	-53.04	232.99	286.03	2.97	48.73

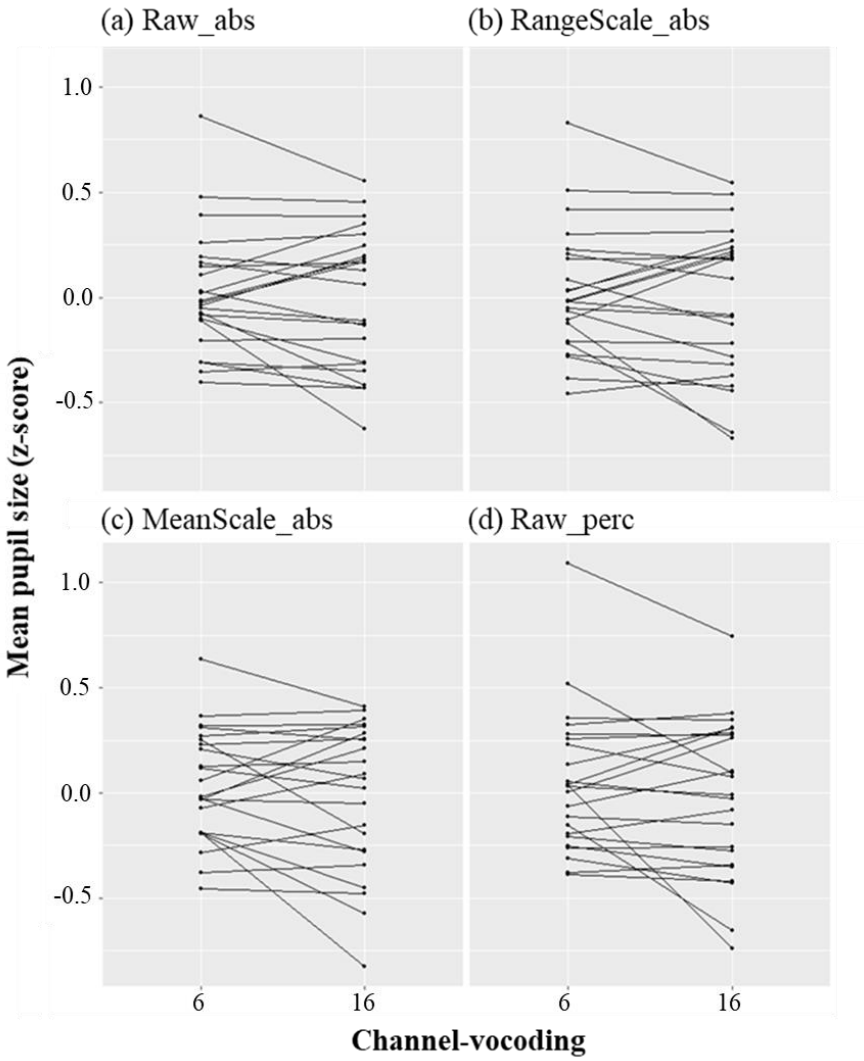

Figure 3. Individuals' mean pupil size values by vocoding condition for the LME/ Bayes data aggregation. All values z-scored to facilitate comparison on a common scale.

Statistical model

Across all statistical models, group data showed pupil size was larger for 6-channel compared to 16-channel vocoded conditions, and to a lesser extent, pupil size was larger for 50% SRT compared to 80% SRT conditions. Model summary statistics, including effect sizes are reported in Table 5. Note the higher standard error for the rmANOVA models, and resulting wider confidence intervals. This is a consequence of aggregating the trials, which increases the apparent variability.

Raw data and absolute change from baseline

Channel-vocoding revealed significant differences between 6 and 16 channel-vocoded sentences, as indicated by 95% CIs that do not exclude zero, for both the LME 95% CI [4.46, 32.19], and Bayes model, 95% CI [3.863, 31.032]. Pupil size was 18.325 and 17.454 units higher in the 6-channel vocoded condition for LME and Bayes models, respectively. The rmANOVA model did not detect a significant difference between the vocoding conditions, although the direction showed the pupil diameter was larger in the 6-channel condition. There were non-significant differences between the SRT conditions across all statistical models.

Range scaled data and absolute change from baseline

Similar to the raw data with absolute baseline correction, channel-vocoding revealed significant differences between 6 and 16 channel-vocoded sentences, as indicated by 95% CIs that do not exclude zero, for both the LME 95% CI [0.001, 0.032], and Bayes model, 95% CI [0.001, 0.032]. Pupil size was 0.017 units higher in the 6-channel vocoded condition for both LME and Bayes models. The rmANOVA model did not detect a significant difference between the vocoding conditions, although the direction of change was as expected. There were non-significant differences between the SRT conditions across all statistical models.

Mean scaled data and absolute change from baseline

Mean-scaled data with absolute baseline correction showed non-significant differences for both vocoding and SRT conditions across all statistical models.

Raw data and percent change from baseline

Raw data with relative percent change from baseline showed non-significant differences for both vocoding and SRT conditions across all statistical models.

Table 5. rmANOVA, LME and Bayes statistical model output for a single pupil size dataset, grouped by data processing method and factors.

Data processing	Factor	Model	Estimate	95% CI	Hedges g
1. Raw and absolute	Channel-vocoding	rmANOVA	16.776	[-8.398, 41.95]	0.179
		LME	18.325	[4.46, 32.19*]	0.073
		Bayes	17.454	[3.863, 31.032*]	0.068
	SRT	rmANOVA	1.18	[-18.936, 16.576]	0.012
		LME	0.957	[-12.911, 14.826]	0.004
		Bayes	0.91	[-12.682, 14.491]	0.004
2. Range scaled and absolute	Channel-vocoding	rmANOVA	0.016	[-0.043, 0.011]	0.17
		LME	0.017	[0.001, 0.032*]	0.06
		Bayes	0.017	[0.001, 0.032*]	0.059
	SRT	rmANOVA	0.007	[-0.025, 0.011]	0.072
		LME	0.007	[-0.009, 0.023]	0.025
		Bayes	0.007	[-0.009, 0.023]	0.025
3. Mean scaled and absolute	Channel-vocoding	rmANOVA	0.004	[-0.014, 0.006]	0.108
		LME	0.005	[-0.001, 0.011]	0.047
		Bayes	0.005	[-0.001, 0.011]	0.046
	SRT	rmANOVA	0.002	[-0.01, 0.006]	0.053
		LME	0.002	[-0.004, 0.008]	0.019
		Bayes	0.002	[-0.004, 0.008]	0.017
4. Raw and percent	Channel-vocoding	rmANOVA	0.399	[-1.361, 0.563]	0.1
		LME	0.489	[-0.191, 1.168]	0.04
		Bayes	0.489	[-0.191, 1.168]	0.039
	SRT	rmANOVA	0.323	[-1.109, 0.463]	0.078
		LME	0.281	[-0.398, 0.96]	0.023
		Bayes	0.352	[-0.188, 0.89]	0.029

6.6 Discussion

The aim of this study was to examine the effect of different scaling methods, baseline corrections, and statistical approaches, on the results and interpretation of the same pupillometric dataset in order to enhance replicability and reliability of published studies in the P-LE literature. This study demonstrated that data processing methods such as scaling, baseline correction, and data aggregation for input to statistical models greatly affects the outcome, as does choice of statistical approach.

The raw and scaled data produced vastly different distributions, as did the type of baseline correction when comparing the raw data. Additionally, aggregating the data for the different statistical models also greatly altered the distributions between the datasets, with the exception of range scaling which appeared relatively robust against aggregation likely due to this method limiting outliers compared to other models. Type of scaling method and baseline correction led to different statistical results and subsequent interpretations of the outcome.

Although, on average, pupil size was larger in 6-channel vocoding and 50% SRT conditions (the expected direction) in the rmANOVA models across data processing types, there were no statistically significant differences between the conditions leading to the interpretation that changes in spectral resolution (channel-vocoding) and SRT did not significantly impact pupil size/ listening effort. The output of the LME and Bayes models demonstrated that pupil size was significantly larger in the 6-channel compared to 16-channel vocoding condition for both the raw and range scaled absolute change from baseline data, suggesting that spectral resolution, in this dataset, influences pupil size/ the amount of effort required to listen to a degraded signal. The pupil response was not significantly affected by changes in SRT across any data processing method or statistical model.

While the raw and range scaled absolute change from baseline showed that channel-vocoding significantly influenced pupil diameter, mean-scaled data with absolute change from baseline and raw data with percent change from baseline did not produce statistically significant results in any statistical model (although the direction of change was as expected). This would lead to the statistical conclusion that neither channel-vocoding nor SRT impact pupil diameter, and therefore listening effort was comparable across conditions. In regards to the mean-scaled data, this is not entirely surprising as Kuchinsky and colleagues (2013) compared model outputs from a rmANOVA and Growth Curve Analysis (GCA) model (discussion forthcoming) demonstrating that the rmANOVA model was insensitive to changes in task parameters on pupil diameter whilst statistically significant differences were observed in the GCA model. A direct comparison is however not possible with the current dataset because Kuchinsky and colleagues used maximum pupil diameter, while the current dataset used mean pupil diameter.

As evidenced from the descriptive statistics (Table 4) and Figures 1 and 2, each processing or aggregation method applied to the dataset led to a non-normal distribution. Further to this, testing for normality of the residuals was not conducted in this study, nor was testing any statistical assumptions. It is commonplace in the P-LE literature to not disclose whether statistical assumptions were met and discussion concerning outliers is sparse. This is also common across disciplines, with empirical research showing that less than 8.5% of the included psychology papers in their study reported testing statistical assumptions and examining outliers (Osborne, Christianson, & Gunter, 2001). The impact of outliers on statistical tests can be detrimental because they can increase the error variance (Osborne & Overbay, 2004), where outlying points will have an excessive effect on the estimates. This is evident in the current dataset when comparing the raw absolute change and percent change from baseline processing method. The ratio of the standard error to the estimate in the raw

absolute change was 0.385 compared to the relative percent change being 0.707 for channel-vocoding/ LME. It was also previously introduced that percent change from baseline calculations can skew distributions and decrease statistical power (Vickers, 2001) which was confirmed in the current set of results. That is, outliers can degrade statistical power (Osborne & Overbay, 2004).

Working with percent change from baseline proves useful when examining which observations may not be legitimate as the values are more easily interpretable than when the data has been scaled. This is because scaling data renders it unitless, which makes it difficult to estimate probable size changes. In comparison, percent changes from baseline makes it easier to identify plausible outliers, for example, a 200% change from baseline appears improbable when the average pupil size is 4.5 mm. Transforming the data to fit a normal distribution is a common procedure for managing outliers (e.g., square root or log-transformations), yet require non-negative data which is widespread in pupillometric datasets. Interpreting results of transformed data also becomes difficult, especially when already-scaled data undergo subsequent transformation.

Removing legitimate outliers can result in unbalanced datasets which is particularly problematic for traditional rmANOVA data structures. These types of data structures are prone to list-wise deletion when observations are missing which can result in sample bias and unequal variance (Abdi, 2010). Conversely, data entered into a LME model retains trial-by-trial observations resulting in long-format data aggregation and does not result in list-wise deletion in the event of missing observations. Assuming the missing observations are random, a LME model is capable of filling in the missing data provided that it can be predicted from the available data. For a comparison of a rmANOVA and LME model using incomplete datasets, see Krueger and Tian (2004).

LME and Bayes models are also reasonably robust to violations of assumptions that typically affect rmANOVAs (Woltman, Feldstain, MacKay, & Rocchi, 2012). A Bayesian approach can also incorporate prior knowledge ('priors') about a distribution into a model which can reduce variance and produce better estimates. Non-Bayesian models on the other hand implicitly use a flat (uniform) fixed prior. The output of a Bayesian model, 'posterior probability', is calculated using both the input data and priors. Priors can be both informative (e.g., sourced from published studies) and uninformative (e.g., vague knowledge). In the analyses presented here, the posterior distribution was largely determined by the data as informative priors were not included in the model. As such, the results closely mimic those of the LME model output (although slightly different due to the way likelihood functions are calculated). Incorporating informative priors leads to the posterior distribution being a combination of prior probability and the data, with highly informative priors decreasing variance of the posterior distribution. Setting informative priors is particularly relevant when sample sizes are small which may prove useful when recruiting highly specialised populations. Since the P-LE research area is rich in prior information, it is possible to use this information in future analyses and meta-analyses.

Recently, some research groups have begun to use GCA to examine the time-course and curve morphology of the pupil in response to varying levels of listening difficulty (Kuchinsky et al., 2016; Kuchinsky et al., 2014; Kuchinsky et al., 2013; Wagner et al., 2016; Wagner et al., 2016; Winn, Edwards, & Litovsky, 2015). GCA, as applied to pupil data, uses mixed-model regression fitting orthogonal polynomials to model the shape of the pupil response over time (see Mirman, 2016, for a comprehensive review). The current study analysed the mean pupil dilation averaged in time-windows over the levels of participant and factors or trials, disregarding the changes in pupil dilation over the time-course of trials. It is possible

that retaining timing information may result in a more sensitive measure of listening effort, or perhaps, a reduction in variability (see below).

There has also been a recent shift in the statistical literature in relation to modelling random effects structures (for alternate discussions see Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015). For example, the LME analyses in this chapter used a random intercept model, however an alternative analysis using both random intercepts and random slopes could have also been applied. This type of analysis would permit the independent variables (SRT and channel-vocoding) to vary for each participant (instead of at group level) which may result in a better model fit, and additionally, explain some of the variability as discussed below (likewise, this type of model structure can be applied to a Bayesian analysis (Sorensen & Vasishth, 2015)). Although this type of analysis has not been applied in the P-LE literature, it is important that future studies consider how various random effects structures influence the results and subsequent interpretations of the results.

Further to this, examining the additional peak-picked parameters such as maximum pupil size and latency to maximum pupil size (as opposed to the mean pupil size used in all analyses here) may provide additional information and prove to be a more sensitive measure of listening effort. At the very least, these metrics should be explored to verify whether they are more robust to individual variability. Extending the comparisons presented in this paper to include GCA and random slopes models and the analysis of maximum and latency to maximum pupil size would have brought additional variability to the interpretation of the results. To contain the paper and avoid redundancy, the authors chose to limit the processing and analyses to the most commonly reported methods, as a demonstration of how different methods apply to a single dataset affect results. The full dataset, however, can be found on the repository for further analysis.

The main comparisons presented in this paper related to group results and showed similar trends despite different results. An additional consideration that appears overlooked in the literature relates to individual variability, as highlighted in Figure 3 across vocoding conditions. For some individuals, the pupil diameter was larger in the 6-channel condition compared to the 16-channel condition, whilst for others, pupil diameter was comparable between conditions, or larger in the 16-channel condition. Whilst error terms or random intercepts were included in all statistical models, the conducting and reporting of group level analyses limits discussion of the pupil's behaviour at the individual level. Whereas this may not be problematic for using pupillometry as a research tool to assess group trends, it brings to light one of the many challenges of using pupillometry as a clinical tool to assess listening effort. That is, if pupillometry is to be considered a tool to assess listening effort in a clinical setting, for example when comparing different programs in hearing aids, the pupil must predictably respond to changes in task difficulty at the level of the individual. Whether the source of interindividual variability stems from the design of the current study, the extent to which an individual's cognitive ability and motivational levels modulated listening effort, that variability in pupil dilation is a general phenomenon, or a combination of these, is unknown. Moving forward, it will be important to consider the extent to which these factors may influence listening effort, and design more robust studies to identify and limit individual variability.

6.7 Conclusion

Further work is required to establish the viability of pupillometric measures to index listening effort in both research and clinical settings. This paper began to address some of the potential issues of reliability and replicability in the research domain, notably, through examining how disparate scaling and baseline corrected data entered into different statistical models can considerably influence the results and discussion of a study's outcome. At the very least,

research output would be strengthened by: 1) standardising data processing and analyses approaches across publications, 2) disclosing test assumptions/ outliers and including histograms, 3) reporting effect sizes/ confidence intervals to facilitate comparisons between studies and aid in future meta-analyses, and 4) making data publicly available to facilitate direct comparisons within- and between studies. Assessing whether pupillometry is a viable clinical tool to evaluate listening effort would be strengthened through: a) disclosing individual variability, b) examining a range of covariates such as cognitive and motivational factors and designing studies to exploit these parameters to determine if they result in decreased variability, and c) detecting and limiting measurement error through more stringent pre-processing methods. Despite its exploratory nature, this study offers insight into some of the methodological issues that warrant consideration when analysing pupillometric datasets.

6.8 References

Abdi, H. (2010). The Greenhouse-Geisser correction. In Salkind N (Ed), *Encyclopedia of Research Design* (pp. 544-548). Thousand Oaks: Sage.

Allen, C. P., Dunkley, B. T., Muthukumaraswamy, S. D., Edden, R., Evans, C. J., Sumner, P., . . . Chambers, C. D. (2014). Enhanced awareness followed reversible inhibition of human visual cortex: a combined TMS, MRS and MEG study. *PLoS ONE, 9*(6), e100350.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software, 20*, 1-37.

Cavanagh, J. F., Wiecki, T. V., Kochar, A., & Frank, M. J. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology: General, 143*(4), 1476.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilatation. *International Journal of Audiology, 36*(3), 155-164.

Krueger, C., & Tian, L. (2004). A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biological Research for Nursing, 6*(2), 151-157.

Kuchinsky, S. E., Vaden Jr, K. I., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2016). Task-related vigilance during word recognition in noise for older adults with hearing loss. *Experimental Aging Research, 42*(1), 50-66.

Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2014). Speech-perception training for older adults with hearing loss impacts word recognition and effort. *Psychophysiology, 51*(10), 1046-1057.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*(1), 23-34.

Lemercier, A., Guillot, G., Courcoux, P., Garrel, C., Baccino, T., & Schlich, P. (2014). Pupillometry of taste: Methodological guide–from acquisition to data processing-and toolbox for MATLAB. *The Quantitative Methods for Psychology, 10*(2), 179-195.

Lenth, R. V., & Hervé, M. (2013). lsmeans: Least-squares means. *R package version, 2*, 11.

Luck, S. J. (2014). *An introduction to the event-related potential technique*: MIT press.

Martins, T. G., Simpson, D., Lindgren, F., & Rue, H. (2013). Bayesian computing with INLA: new features. *Computational Statistics & Data Analysis, 67*, 68-83.

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., & Lyxell, B. (2017). Objective assessment of listening effort: co-registration of pupillometry and EEG. *Trends in Hearing*.

Mirman, D. (2016). *Growth curve analysis and visualization using R*: CRC Press.

Osborne, J. W., Christianson, W. R., & Gunter, J. S. (2001). *Educational psychology from a statistician's perspective: A review of the quantitative quality of our field.* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation, 9*(6), 1-12.

Pinheiro, J., Bates, D., DebRoy, S., & Sarkar, D. (2014). R Core Team (2014) nlme: linear and nonlinear mixed effects models. R package version 3.1-117. *Available at h ttp://CRAN. R-project. org/package= nlme*.

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling.* Paper presented at the Proceedings of the 3rd International Workshop on Distributed Statistical Computing.

Rue, H., Martino, S., Lindgren, F., Simpson, D., Riebler, A., & Krainski, E. T. (2014). INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximaxion. *R package version 0.0-1404466487, URL* http://www. *R-INLA.org*.

Shakespeare, T. P., Gebski, V. J., Veness, M. J., & Simes, J. (2001). Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *The Lancet, 357*(9265), 1349-1353.

Siegle, G. J., Ichikawa, N., & Steinhauer, S. (2008). Blink before and after you think: blinks occur prior to and following cognitive load indexed by pupillary responses. *Psychophysiology, 45*(5), 679-687.

Singmann, H., Bolker, B., & Westfall, J. (2015). afex: Analysis of Factorial Experiments. *R package version 0.13-145.*

Sorensen, T., & Vasishth, S. (2015). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *arXiv preprint arXiv:1506.06201*.

R Core Team. (2008). R: A language and environment for statistical computing.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K., & van der Werf, M. J. (2006). Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics, 7*(1), 1.

Vickers, A. J. (2001). The use of percentage change from baseline as an outcome in a controlled trial is statistically inefficient: a simulation study. *BMC Medical Research Methodology, 1*(1), 1.

Wagner, A., Pals, C., de Blecourt, C. M., Sarampalis, A., & Başkent, D. (2016). Does Signal Degradation Affect Top–Down Processing of Speech? In P. van Dijk, D. Başkent, E.

Gaudrain, E. de Kleine, A. Wagner, & C. Lanting (Eds.), Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing (pp. 297-306). Cham: Springer International Publishing.

Wagner, A. E., Toffanin, P., & Başkent, D. (2016). The timing and effort of lexical access in natural and degraded speech. *Frontiers in Psychology*, 7: 398.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, 7: 345.

Wickham, H. (2017). tidyverse: Easily install and load 'tidyverse'packages [Software].

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing, 36*(4), e153-e165.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing, 31*(4), 480-490.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: neural correlates of pupil size as a measure of cognitive listening load. *Neuroimage, 101*, 76-86.

Chapter 7 General discussion

Listening effort is a multifaceted construct, and recent research is only now shedding light on its complexities. This thesis explored some of the listener-internal and external factors which modulate listening effort. Specifically, listener-internal factors were investigated by evaluating how working memory capacity interacts with subjective and physiological measures of listening effort, whereas listener-external factors were evaluated by examining the impact of spectral resolution (channel-vocoding) and performance parameters on subjective and physiological measures of listening effort. We also assessed how data processing strategies and statistical approaches can affect the interpretation of physiological measures. These studies were conducted within the overarching context of exploring the potential for a physiological measure of listening effort to be useful in a clinical environment. We conducted a study to address four research questions, each investigating different aspects relevant to listening effort measures, with the same sample of young adults with normal hearing.

Together, the results presented in this thesis will inform current international research in the development of clinically viable measures of listening effort, where promising group trends are accumulating. This is particularly true for important research questions related to physiological measures of listening effort, such as:

- Does working memory capacity and different types of spectral resolution (channel-vocoding), similar to the signal provided by a cochlear implant, affect responses measured during effortful listening?
- How do the different physiological measures respond when modulating spectral resolution and background noise, and do they index the same construct?
- Are current published results which are related to physiological measures of listening effort sufficiently consistent and interpretable to be used clinically?

1) The contribution of working memory on measures of effortful listening

Our findings indicated that there was a complex interaction between working memory capacity and self-rated perceptions of listening effort on SRT and channel-vocoding (Chapter 3), but not with the physiological responses measured in Chapter 5.

Chapter 3 investigated whether working memory capacity influenced perceived effort ratings and included two parts. The main result of Study (a) demonstrated that those individuals with higher working memory capacity, on average, rated listening effort differently across the levels of speech reception thresholds and spectral resolution (16- and 6-channel vocoded sentences). In contrast, individuals with lower working memory capacity, on average, did not rate listening effort differently when spectral resolution was greatest. There are two potential explanations for this. First, individuals with higher working memory may be better able to maintain their attention on a task (or be less distracted by noise) than those who score lower on the test. That is, working memory capacity may be related to greater suppression of background noise interference enabling individuals to make a perceived effort judgement on the quality of the signal rather than the level of background noise. Second, individuals with higher working memory capacity might be better equipped to more rapidly adapt to acoustic degradation, and therefore quickly learn the acoustic patterns in each condition permitting them to be able to invest more effort during the more difficult task conditions. While this hypothesis was not directly tested in the current thesis, it is in line with studies demonstrating that when hearing is impaired, individuals with higher working memory capacity more quickly adapt to new hearing aid signal-processing algorithms (Lunner, 2003; Ng, Rudner, Lunner, Pedersen, & Rönnberg, 2013; Rudner, Foo, Rönnberg, & Lunner, 2009). While this same finding was not statistically significant in Study (b) (Chapter 3) where a similar trend between participants with high and low working memory capacity was found, performance appeared to influence ratings particularly for those with lower working memory capacity.

We further assessed whether measures of working memory capacity interacted with the physiological measures of listening effort (Chapter 5). The same measure of working memory capacity used in Chapter 3 was used to model the interaction in this study using the participants and the physiological data presented in Chapter 4. The results indicated that there were no significant interactions between working memory capacity, performance levels, and/ or spectral resolution, for either the pupil dilation or alpha power change. While this observation may indicate that working memory capacity does not play a role in the pupil or alpha responses to spectral resolution and performance levels, this finding should be interpreted with caution; working memory capacity was assessed during a separate testing session to the physiological testing session and was assessed offline.

Pichora-Fuller and colleagues (2016) suggested that cognitive assessments, such as a measure of working memory capacity, may be useful to index listening effort. While the current results suggest that differences in working memory capacity do interact with perceived listening effort, consistent with the results from Rudner et al. (2012) obtained in older hearing-impaired individuals, this was not the case for the physiological measures. On the other hand, Wendt et al. (2016) found a positive correlation between pupil size during an online measure of working memory capacity and a sentence-in-noise task. Therefore, it would be of interest to explore whether an interaction does exist between an online measure of working memory and pupil dilation and alpha power with the current study protocol. This is important for future research into the clinical applicability of physiological measures of listening effort. That is, the more robust a measure is to listener-internal differences, such as individual differences in the allocation of cognitive resources, the more predictive power it will have as a clinical assessment.

While working memory capacity has been a focus of this thesis, both attention and speed of phonological processing have also been suggested as potentially indexing listening effort

(Pichora-Fuller et al., 2016). Like working memory, attention is also limited in capacity (Raymond, Shapiro, & Arnell, 1992). Allocation of attentional resources is critical for listening to speech in noise and was described by Cherry (1953) as the ability to focus attention on a speech signal in complex listening environments (i.e., 'the cocktail party effect'). Processing an incoming auditory signal such as conversational speech also requires the rapid integration/matching of acoustic cues to phonological and semantic representations stored in memory. Processing speed may therefore constrain other cognitive operations. For example, when it becomes slowed, the ability to focus attention or update items in working memory becomes challenged (discussed in greater detail in (3) below). More research is needed to better understand how these cognitive operations interact with physiological measures of listening effort.

2) Effects of spectral resolution and background noise on measures of listening effort

This thesis also investigated how spectral resolution and performance levels influence subjective (Chapter 3) and physiological (Chapter 4) measures of listening effort.

The results of Chapter 3 demonstrated that on average, individuals rated perceived listening effort higher in the more challenging condition (e.g., 50% SRT compared with 80% SRT). The results demonstrated that on average, individuals rated perceived listening effort higher in the more challenging condition, but the ratings were not influenced by spectral resolution. That is, perceived listening effort ratings appear to be sensitive to performance levels, with individuals likely rating effort higher in the 50% SRT condition due to greater task difficulty and recognised inferior performance compared to the 80% SRT condition. This finding supports the notion that individuals may be inclined to rate their estimated performance on a task as opposed to rating the effort that was invested in performing the task (McGarrigle et al., 2014).

In contrast to this, our results indicated that the physiological measures - pupil dilation and alpha power change - were sensitive to changes in spectral resolution (Chapter 4). On average, pupil dilation was larger when spectral resolution was lowest (6-channel vocoding), while alpha power change was greatest with higher spectral resolution (16-channel vocoded sentences). The pupil size change, but not alpha power change, was additionally sensitive to changes in performance levels, significantly decreasing in diameter as performance levels increased towards ceiling. Further, the pupil response was sensitive to changes in task accuracy; when sentences were only partially recalled, the pupil diameter was greater compared to when sentences were correctly recalled.

The present results are significant in at least two major respects. Firstly, the findings suggest that individuals are inclined to rate their estimated performance, or success with doing the task, not the perceived effort that was invested to perform the task. The only metric which is commonly used with standard clinical speech recognition tests is an individual's performance level. As such, an assessment of perceived listening effort ratings during a task is not likely to increase the sensitivity, or complement, current clinical assessments. The second major finding is that pupil dilation appears to be a more sensitive measure of listening effort than alpha power change. The pupil response was responsive to a variety of factors that contribute to effortful listening, including performance and task accuracy, whereas alpha power change only significantly responded to changes in spectral resolution. In terms of clinical viability, the pupil response may therefore be a more sensitive measure of listening effort. However, it could alternately be argued that the pupil's responsiveness to a variety of factors is, in fact, a limitation of its clinical viability as its sensitivity is too broad. Alpha power change was only associated with changes in spectral resolution and was robust against performance differences and task accuracy. If changes in spectral resolution predictably equate to both increased and decreased listening effort across a more challenging and less challenging listening task,

respectively, then its insensitivity to other parameters may be a strength when considering its clinical application.

3) Comparing pupil dilation and alpha oscillations within the construct of listening effort

Our results indicated that during an effortful listening task, the changes measured in pupil diameter and alpha oscillations are not correlated (Chapter 5). The rationale for this study was that if each measure, as claimed, indexes listening effort, and that listening effort is a single construct, then factor manipulations should have led to similar responses in both measures. After accounting for unequal trials across the measures (e.g., where one pupil trial may have been rejected due to blink contamination, the corresponding alpha trial was also removed from the analysis), the results showed that the two measures were not correlated, a finding that has been previously demonstrated during both a listening task (McMahon et al., 2016) and in the reading domain (Scharinger, Kammerer, & Gerjets, 2015). A possible explanation for this might be that each measure is assessing a different aspect of listening effort, which is inherently multifaceted rather than a single dimensional construct. Another explanation for this is that the measures may reflect differences in the types of strategies applied by each individual. For example, it is unclear how differences in motivational strategies between individuals influence measures of listening effort. Some individuals may be highly motivated to maintain their performance across the testing session, whereas others may be less inclined. Changes in the pupil response have indeed been associated with motivation, as manipulated through performance-based monetary reward (Heitz et al., 2008) and goal-orientation (Gilzenrat, Cohen, Rajkowski, & Aston-Jones, 2003; see Aston-Jones & Cohen, 2005, for review). Gilzenrat et al. (2010) also showed that a larger pupil size in the baseline region was related to slower reaction times (i.e., speed of processing), increased variability, decreased performance accuracy, and task disengagement. As such, the responsiveness of the pupil appears to be modulated by both bottom-up sensory information

(e.g., acoustic degradation), and top-down information (e.g., motivation), through cortical locus coeruleus feedback loops (Aston-Jones & Cohen, 2005; Sarter, Gehring, & Kozak, 2006). While changes in alpha power, particularly in the parietal region, have been implicated in attentional processes (Benedek, Schickel, Jauk, Fink, & Neubauer, 2014), the neurophysiological mechanisms underlying alpha oscillations are not well understood. It is possible that the responses are partly interrelated but activated differently depending on both levels of arousal and task difficulty. Further research is needed to better understand how different factors of attention, arousal and motivation might influence the different physiological measures.

There are a range of differing EEG and pupillometry metrics that may be used to assess listening effort and it is possible that more parallel measures may explain the lack of correlation in this chapter. For example, the analysis in Chapter 4 was conducted using mean power in the oscillatory band, and maximum pupil size. It could be argued that a measure of maximum alpha power may provide a more parallel process when compared to maximum pupil size (or mean alpha power compared to mean pupil size), and this presents a limitation in the current thesis.

In regards to EEG, previous research (notably Obleser et al., 2012, on which most of this thesis' foundation has been based) used mean alpha power as a measure of listening effort. Obleser and colleagues have consistently found that mean alpha power and working memory capacity are additive, suggesting that they may have a common oscillatory network. As such, mean alpha power was the metric selected for analysis within this thesis. Despite this, maximum alpha power could also be an interesting parameter to explore, yet what a 'maximum oscillation' reflects requires further debate, and also discussion is warranted regarding baseline correcting maximum oscillatory power. For example, if percent baseline

correction is applied to maximum alpha power, the data will likely be highly skewed and may require transformation (see discussion in Chapter 6).

On the other hand, a range of pupil metrics have been explored, ranging from mean and maximum pupil dilation, through to latency to maximum pupil size, blink latency, and saccades. In this thesis, one of the main questions of Chapter 5 was whether maximum pupil size was related to working memory capacity and this line of questioning had only been investigated using maximum pupil size at the time this experiment was conceived (Koelwijn, 2012; Zekveld & Kramer, 2014). We therefore chose to explore whether a relationship existed between mean alpha power and maximum pupil dilation based on the previous literature, however further exploration across all EEG and pupillometry metrics is warranted.

4) Methodological considerations when interpreting physiological measures of listening effort

This thesis also examined how different data processing and statistical analyses used in the literature related to physiological measures of listening effort influence the results and interpretation of a single dataset (Chapter 6). Using our pupillometric data as an illustration, the results demonstrated that the varying processing and analyses methods do, in fact, greatly alter the results and their interpretation. This has consequences for understanding how different factors and population demographics interact with the pupil response across studies, leading to disparate conclusions regarding what parameters influence listening effort. This has further implications for the design of future studies.

Perhaps most importantly, this study highlighted the large amount of individual variability in the data. The breakdown of individual results across vocoding conditions demonstrated that the pupil response greatly differs between individuals. Some participants had much larger pupil diameters in the 6-channel condition compared to 16-channel sentences (the expected

direction), yet others differed little between conditions, and some individuals showed the opposite pattern. That is, inter-individual pupil diameters across factors are highly variable. Intraclass correlation coefficients were conducted to assess intraindividual reliability in the pupil data (Chapter 4), and the results of the analysis determined that there was a very high degree of intraindividual reliability (i.e., across factors, individuals' pupil changes respond highly consistently). However, if pupillometry is to be considered a viable tool to assess listening effort in a clinical setting, it must respond predictably and reliably at the individual level across factor manipulations. While results indicate that the pupil does reliably respond at the group level, further analyses revealed it does not predictably respond to the manipulation of factors at the individual level.

Chapter 6 discussed how some statistical models may be more sensitive to detecting changes in listening effort and may also lead to a reduction in individual variability. For example, various random effects structures were introduced, however these analyses were outside of the scope of the chapter as it was focussed on comparing what models have already been applied in the literature. However, as described above, it is important that future studies examine how the different random effects structures influence the datasets, as this may result in better fitting models, and also account for some of the individual variability evident when using traditional repeated measures ANOVAs and random intercept models.

Chapter 8 Implications for future studies

More research is needed to understand the mechanisms underpinning listening effort and how they contribute to individual variability if a clinical tool is to be developed and implemented. A greater focus on experimental design may produce findings that account for the large amount of individual variability in the present thesis. This may include more closely examining the mechanisms that comprise listening effort, such as motivation, and/ or experimental design more generally.

The recently published Framework for Understanding Effortful Listening (Pichora-Fuller et al., 2016; Figure 1), along with the Strauss & Francis' (2017) model of attention and effortful listening both highlight the importance of motivation in understanding listening effort, with recent studies elucidating the relative importance of motivation during listening tasks. For example, Richter (2016) showed that when listening in more challenging conditions, cardiovascular reactivity is higher when monetary compensation for successful performance is greatest, and low when there is no monetary compensation. This suggests that task demands alone cannot account for changes in listening effort. A similar effort-reward trade-off has been demonstrated in pupillometry studies resulting in an inverted u-shaped pattern capturing the multidimensionality of listening effort (Gilzenrat et al., 2010). In our series of studies, the motivational dimension of listening effort was not manipulated and participants were therefore free to mobilise motivation without constraint. For example, when task difficulty was high, only a proportion of participants may have been motivated to correctly recall the sentences. Some participants, on the other hand, may have disengaged in the task, resulting in low motivational input. Not controlling the motivational dimension of listening effort may therefore explain some of the variability found in the current thesis.

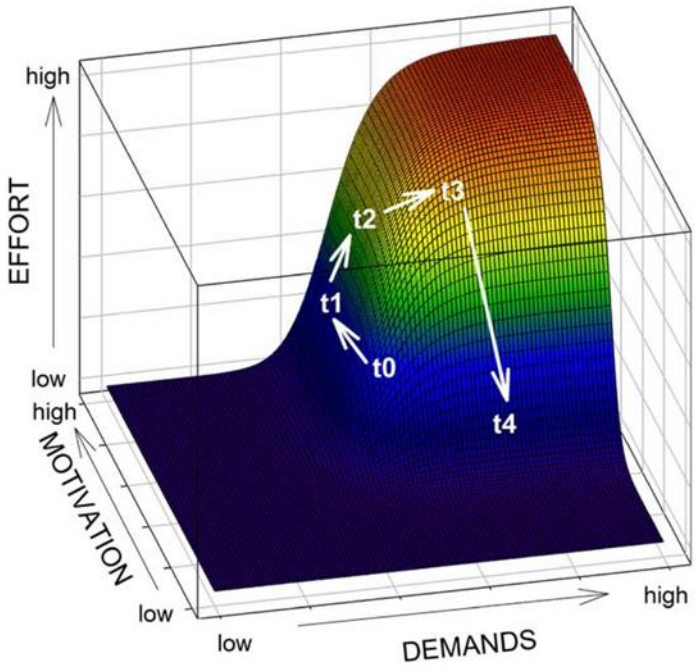

Figure 1. The 3-D representation of the FUEL illustrating the three-way relationship between listening effort, motivation and task demands from Pichora-Fuller et al. (2016). t0-t4 refer to time, discussed in more detail in Pichora-Fuller et al. (2016).

A greater focus on experimental design may also produce findings which minimise individual variability. This may comprise disentangling the cognitive and physiological mechanisms involved in listening effort such as motivation (as above) or exploring how different random effects structures in statistical design influence datasets and the resultant interpretations, and examine whether this reduces variability. Future studies may also want to consider how stimuli may affect studies of listening effort and individual variability. For example, using predictable sentences (such as in this thesis), permitted individuals to use sentential context to fill in the gaps when some words were masked by noise. Variability may therefore arise due to individuals diverging in their crystallised knowledge (e.g., verbal ability). Using stimuli such as CNC wordlists may partly inhibit these processes and result in reduced variability between individuals. Yet developing experimental studies to reduce individual variability may result in scenarios far removed from what is encountered by individuals in the real-

world, and therefore what may be clinically relevant. It is infrequent that engaging in conversation in the real-world requires listening to serially listed CNC words in fluctuating SNRs. As such, there must exist a trade-off between what constitutes robust experimental design, and the generalisability of results to the real-world. Whilst a Patient-Reported Outcome Measure (PROM) of perceived listening effort is in development (Hughes, Rapport, Boisvert, McMahon, & Hutchings, 2017), at present, work is needed to establish the link between the listening effort exerted in experimental studies (that may be clinically relevant), to that of the real-world.

It is also important that future research consider alternate physiological tools for assessment. This thesis explored pupil dilation and alpha power change as potential clinical tools due to breadth of literature, and clinical relevancy such as their non-invasiveness, portability, and user-friendliness. However other tools such as cardiovascular measures or skin conductance may show reduced variability and prove to be a more robust measure of listening effort for a clinical setting. Future studies examining alternate measures should therefore consider disclosing variation in participant responses to establish insight into this matter.

Further, this thesis used mean and maximum pupil size and mean alpha power change as potential indicators of listening effort. However there are multiple parameters in both pupillometry and the EEG that warrant further investigation. For example, latency to maximum pupil size, blink latency and saccades may prove more reliable than the mean and maximum pupil changes reported here. There are also multiple dimensions in the EEG (both oscillatory and event-related potentials) that could be exploited to further assess listening effort. For example, one study has shown that power change in the theta oscillatory band is linked to changes in listening effort (Wisniewski et al., 2015). Similarly, a relationship between listening effort and both the P300 and late positive potential (LPP) have recently been demonstrated (Bertoli & Bodmer, 2014). More research is required to investigate

whether these parameters respond more predictably and reliably both within and between individuals than those reported in this thesis.

8.1 Conclusion

This thesis builds on the remarkable progress the field has made in moving towards a physiological measure of listening effort. Examining both subjective and physiological measures of listening effort highlighted key differences between how each respond to changes in task difficulty. Importantly, subjective ratings of listening effort were associated with performance levels and not changes in spectral resolution; only when working memory capacity is taken into account, do differences in perceived effort ratings emerge. That individuals rate performance levels suggests that a measure of perceived listening effort would not enhance current clinical assessments, and that a physiological measure may be more suited to increasing assessment sensitivity. While both physiological measures assessed in this thesis appeared to be sensitive to the manipulations of task factors (e.g., pupil dilation: performance levels, spectral resolution, and task accuracy; alpha power: spectral resolution), the measures were not correlated suggesting each may be indexing a different aspect of listening effort. However, precisely which aspect of listening effort that is best assessed to predict specific-task performance or real-world difficulties remains an open question. What emerges from this thesis is that moving towards a physiological measure of listening effort for clinical implementation requires careful methodological consideration in the design and analysis of studies with a focus on limiting individual variability. Continued research into the mechanisms underpinning listening effort, and their physiological correlates, will greatly contribute towards increasing the sensitivity of standard audiological assessments.

References

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing, 38*(1), e39-e48.

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47*(sup2), S53-S71.

Andersen, P., Andersson, S., & Lomo, T. (1968). Thalamo-cortical relations during spontaneous barbiturate spindles. *Electroencephalography and Clinical Neurophysiology, 24*(1), 90-90.

Arlinger, S., Lunner, T., Lyxell, B., & Pichora-Fuller, K. M. (2009). The emergence of cognitive hearing science. *Scandinavian Journal of Psychology, 50*(5), 371-384.

Aston-Jones, G., & Bloom, F. (1981). Nonrepinephrine-containing locus coeruleus neurons in behaving rats exhibit pronounced responses to non-noxious environmental stimuli. *Journal of Neuroscience, 1*(8), 887-900.

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *Journal of Neuroscience, 14*(7), 4467-4480.

Aston-Jones, G., Rajkowski, J., & Kubiak, P. (1997). Conditioned responses of monkey locus coeruleus neurons anticipate acquisition of discriminative behavior in a vigilance task. *Neuroscience, 80*(3), 697-715.

Aston-Jones, G., Rajkowski, J., & Cohen, J. (1999). Role of locus coeruleus in attention and behavioral flexibility. *Biological Psychiatry, 46*(9), 1309-1320.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience, 28*, 403-450.

Başar, E., Schürmann, M., Başar-Eroglu, C., & Karakaş, S. (1997). Alpha oscillations in brain functioning: an integrative theory. *International Journal of Psychophysiology, 26*(1), 5-29.

Başar, E., Başar-Eroglu, C., Karakaş, S., & Schürmann, M. (2001). Gamma, alpha, delta, and theta oscillations govern cognitive processes. *International Journal of Psychophysiology, 39*(2), 241-248.

Bench, R., & Doyle, J. (1979). *The Bamford-Kowal-Bench/Australian version (BKB/A) Standard Sentence Lists.* Carlton, Victoria: Lincoln Institute.

Benedek, M., Schickel, R. J., Jauk, E., Fink, A., & Neubauer, A. C. (2014). Alpha power increases in right parietal cortex reflects focused internal attention. *Neuropsychologia, 56*, 393-400.

Bertoli, S., & Bodmer, D. (2014). Novel sounds as a psychophysiological measure of listening effort in older listeners with and without hearing loss. *Clinical Neurophysiology, 125*(5), 1030-1041.

Borg, G. A. (1982). Psychophysical bases of perceived exertion. *Medicine & Science in Sports & Exercise, 14*(5), 377-381.

Bouret, S., Duvel, A., Onat, S., & Sara, S. J. (2003). Phasic activation of locus coeruleus neurons by the central nucleus of the amygdala. *Journal of Neuroscience, 23*(8), 3491-3497.

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America, 25*(5), 975-979.

Cohen, M. X. (2017). Where does EEG come from and what does it mean? *Trends in Neurosciences, 40*(4), 208-218.

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience, 3*(3), 201-215.

Corbetta, M., Patel, G., & Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron, 58*(3), 306-324.

Danermark, B., Cieza, A., Gimigliano, F., Granberg, S., Hickson, L., Kramer, S. E., . . . Strömgren, J. P. (2010). International classification of functioning, disability, and health core sets for hearing loss: a discussion paper and invitation. *International Journal of Audiology, 49*(4), 256-262.

Dawes, P., Fortnum, H., Moore, D. R., Emsley, R., Norman, P., Cruickshanks, K., . . . Lutman, M. (2014). Hearing in middle age: a population snapshot of 40–69 year olds in the UK. *Ear and Hearing, 35*(3), e44.

Dorman, M. F., Spahr, A., Gifford, R. H., Cook, S., Zhang, T., Loiselle, L., . . . Schramm, D. (2012). Current research with cochlear implants at Arizona State University. *Journal of the American Academy of Audiology, 23*(6), 385-395.

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing, 34*(3), 261-272.

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *Journal of Speech and Hearing Disorders, 47*(2), 189-193.

Edwards, B. (2007). The future of hearing aid technology. *Trends in Amplification, 11*(1), 31-46.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: ease of listening, word recognition, and attentional effort. *Ear and Hearing, 13*(2), 80-86.

Fraser, S., Gagné, J.P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing

visual speech cues. *Journal of Speech, Language, and Hearing Research, 53*(1), 18-33.

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America, 110*(2), 1150-1163.

Gagné, J.P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: a review. *Trends in Hearing, 21*, 1-25.

Gatehouse, S., & Noble, W. (2004). The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology, 43*(2), 85-99.

Gilzenrat, M. S., Cohen, J. D., Rajkowski, J., & Aston-Jones, G. (2003). Pupil dynamics predict changes in task engagement mediated by locus coeruleus. In *Society for Neuroscience Abstracts* (Vol. 515, p. 19).

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience, 10*(2), 252-269.

Gosselin, P. A., & Gagné, J.P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*(3), 944-958.

Hällgren, M. (2005). *Hearing and cognition in speech comprehension. Methods and applications.* (Doctoral dissertation, Linköping University).

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology, 52*, 139-183.

Heitz, R. P., Schrock, J. C., Payne, T. W., & Engle, R. W. (2008). Effects of incentive on working memory capacity: behavioral and pupillometric data. *Psychophysiology, 45*(1), 119-129.

Herrmann, C. S., Fründ, I., & Lenz, D. (2010). Human gamma-band activity: a review on cognitive and behavioral correlates and network models. *Neuroscience & Biobehavioral Reviews, 34*(7), 981-992.

Hicks, C. B., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research, 45*(3), 573-584.

Hornsby, B. W. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing, 34*(5), 523-534.

Hua, H., Karlsson, J., Widén, S., Möller, C., & Lyxell, B. (2013). Quality of life, effort and disturbance perceived in noise: A comparison between employees with aided hearing impairment and normal hearing. *International Journal of Audiology, 52*(9), 642-649.

Hua, H., Emilsson, M., Kähäri, K., Widén, S., Möller, C., & Lyxell, B. (2014). The impact of different background noises: effects on cognitive performance and perceived disturbance in employees with aided hearing impairment and normal hearing. *Journal of the American Academy of Audiology, 25*(9), 859-868.

Hua, H., Anderzén-Carlsson, A., Widén, S., Möller, C., & Lyxell, B. (2015). Conceptions of working life among employees with mild-moderate aided hearing impairment: a phenomenographic study. *International Journal of Audiology, 54*(11), 873-880.

Hughes, S. E., Rapport, F. L., Boisvert, I., McMahon, C. M., & Hutchings, H. A. (2017). Patient-reported outcome measures (PROMs) for assessing perceived listening effort in hearing loss: protocol for a systematic review. *BMJ Open, 7*(5), e014995.

Humes, L. E. (2007). The contributions of audibility and cognitive factors to the benefit provided by amplified speech to older adults. *Journal of the American Academy of Audiology, 18*(7), 590-603.

Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the alpha band (9–12 Hz) increase with memory load during retention in a short-term memory task. *Cerebral Cortex, 12*(8), 877-882.

Jensen, O., & Mazaheri, A. (2010). Shaping functional architecture by oscillatory alpha activity: gating by inhibition. *Frontiers in Human Neuroscience,* 4: 186.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Englewood Cliffs, NJ: Prentice-Hall.

Kaiser, J., Heidegger, T., Wibral, M., Altmann, C. F., & Lutzenberger, W. (2007). Alpha synchronization during auditory spatial short-term memory. *Neuroreport, 18*(11), 1129-1132.

Karrasch, M., Laine, M., Rapinoja, P., & Krause, C. M. (2004). Effects of normal aging on event-related desynchronization/synchronization during a memory task in humans. *Neuroscience Letters, 366*(1), 18-23.

Kiessling, J., Pichora-Fuller, M. K., Gatehouse, S., Stephens, D., Arlinger, S., Chisolm, T., . . . von Wedel, H. (2003). Candidature for and delivery of audiological services: special needs of older people. *International Journal of Audiology, 42*, Suppl 2, 2s92-101.

Klimesch, W. (1996). Memory processes, brain oscillations and EEG synchronization. *International Journal of Psychophysiology, 24*(1), 61-100.

Klimesch, W., Doppelmayr, M., Pachinger, T., & Ripper, B. (1997). Brain oscillations and human memory: EEG correlates in the upper alpha and theta band. *Neuroscience Letters, 238*(1), 9-12.

Klimesch, W., Doppelmayr, M., Schwaiger, J., Auinger, P., & Winkler, T. (1999). Paradoxical'alpha synchronization in a memory task. *Cognitive Brain Research, 7*(4), 493-501.

Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in Cognitive Sciences, 16*(12), 606-617.

Klink, K., Schulte, M., & Meis, M. (2012). Measuring listening effort in the field of audiology—a literature review of methods (part 2). *Zeitschrift für Audiol, 51*, 60-67.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291-300.

Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hearing Research, 312*, 114-120.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2014). The influence of informational masking on speech perception and pupil response in adults with hearing impairment. *The Journal of the Acoustical Society of America, 135*(3), 1596-1606.

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research, 323*, 81-90.

Koss, M. C. (1986). Pupillary dilation as an index of central nervous system α2-adrenoceptor activation. *Journal of Pharmacological Methods, 15*(1), 1-19.

Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *International Journal of Audiology, 45*(9), 503-512.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*(1), 23-34.

Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2014). Speech-perception training for older adults with hearing loss impacts word recognition and effort. *Psychophysiology, 51*(10), 1046-1057.

Kuchinsky, S. E., Vaden Jr, K. I., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2016). Task-related vigilance during word recognition in noise for older adults with hearing loss. *Experimental Aging Research, 42*(1), 50-66.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry a window to the preconscious? *Perspectives on Psychological Science, 7*(1), 18-27.

Leiberg, S., Lutzenberger, W., & Kaiser, J. (2006). Effects of memory load on cortical oscillatory activity during auditory pattern working memory. *Brain Research, 1120*(1), 131-140.

Lemke, U., & Besser, J. (2016). Cognitive load and listening effort: Concepts and age-related considerations. *Ear and Hearing, 37*, 77S-84S.

Lopes da Silva, F. L., van Lierop, T., Schrijer, C., & Storm van Leeuwen, W. (1973). Organization of thalamic and cortical alpha rhythms: spectra and coherences. *Electroencephalography and Clinical Neurophysiology, 35*(6), 627-639.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *International Journal of Audiology, 42*, S49-S58.

Lunner, T., Rudner, M., & Rönnberg, J. (2009). Cognition and hearing aids. *Scandinavian Journal of Psychology, 50*(5), 395-403.

Luts, H., Eneman, K., Wouters, J., Schulte, M., Vormann, M., Büchler, M., . . . Froehlich, M. (2010). Multicenter evaluation of signal enhancement algorithms for hearing aids). *The Journal of the Acoustical Society of America, 127*(3), 1491-1505.

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology Section A, 58*(1), 22-33.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology, 53*(7), 433-440.

McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., . . . Graham, P. L. (2016). Monitoring alpha oscillations and pupil dilation across the performance-intensity function. *Frontiers in Psychology*, 7: 745.

Mishra, S. (2014). *Exploring Cognitive Spare Capacity: Executive Processing of Degraded Speech.* (Doctoral dissertation, Linköping University).

Ng, E. H. N., Rudner, M., Lunner, T., Pedersen, M. S., & Rönnberg, J. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *International Journal of Audiology, 52*(7), 433-441.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common alpha oscillatory network. *The Journal of Neuroscience, 32*(36), 12376-12383.

Ohlenforst, B., Zekveld, A. A., Jansma, E. P., Wang, Y., Naylor, G., Lorens, A., . . . Kramer, S. E. (2017). Effects of Hearing Impairment and Hearing Aid Amplification on Listening Effort: A Systematic Review. *Ear and Hearing, 38*(3), 267.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia, 9*(1), 97-113.

Pesonen, M., Björnberg, C. H., Hämäläinen, H., & Krause, C. M. (2006). Brain oscillatory 1–30Hz EEG ERD/ERS responses during the different stages of an auditory memory search task. *Neuroscience Letters, 399*(1), 45-50.

Peters, M. L., Godaert, G. L., Ballieux, R. E., van Vliet, M., Willemsen, J. J., Sweep, F. C., & Heijnen, C. J. (1998). Cardiovascular and endocrine responses to experimental stress: effects of mental effort and controllability. *Psychoneuroendocrinology, 23*(1), 1-17.

Petersen, E. B., Wöstmann, M., Obleser, J., Stenfelt, S., & Lunner, T. (2015). Hearing loss impacts neural alpha oscillations under adverse listening conditions. *Frontiers in Psychology,* 6: 177.

Pichora-Fuller, K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America, 97*(1), 593-608.

Pichora-Fuller, K. (2003). Cognitive aging and auditory information processing. *International Journal of Audiology, 42,* 2S26-22S32.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., . . . Mackersie, C. L. (2016). Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing, 37,* 5S-27S.

Picou, E. M., & Ricketts, T. A. (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *International Journal of Audiology, 53*(6), 418-426.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560-569.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology, 20*(3), 241-248.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary suppression of visual processing in an RSVP task: An attentional blink? *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 849.

Richter, M. (2016). The moderating effect of success importance on the relationship between listening demand and listening effort. *Ear and Hearing, 37,* 111S-117S.

Rudner, M., Foo, C., Rönnberg, J., & Lunner, T. (2009). Cognition and aided speech recognition in noise: Specific role for cognitive factors following nine-week experience with adjusted compression settings in hearing aids. *Scandinavian Journal of Psychology, 50*(5), 405-418.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology, 23*(8), 577-589.

Rudner, M. (2016). Cognitive spare capacity as an index of listening effort. *Ear and Hearing, 37*, 69S-76S.

Sarter, M., Gehring, W. J., & Kozak, R. (2006). More attention must be paid: the neurobiology of attentional effort. *Brain Research Reviews, 51*(2), 145-160.

Scharinger, C., Kammerer, Y., & Gerjets, P. (2015). Pupil Dilation and EEG Alpha Frequency Band Power Reveal Load on Executive Functions for Link-Selection Processes during Text Reading. *PLoS ONE, 10*(6), e0130608.

Schneider, B., Pichora-Fuller, M. K., & Daneman, M. (2010). Effects of senescent changes in audition and cognition on spoken language comprehension. In S. Gordon-Salant, D. R. Frisina, A. N. Popper & R. R. Fay (Eds.), *Springer handbook of auditory research: The aging auditory system (vol. 34,)*. New York: Springer

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research, 58*(6), 1781-1792.

Sergeyenko, Y., Lall, K., Liberman, M. C., & Kujawa, S. G. (2013). Age-related cochlear synaptopathy: an early-onset contributor to auditory functional decline. *Journal of Neuroscience, 33*(34), 13686-13694.

Souza, P., & Arehart, K. (2015). Robust relationship between reading span and speech recognition in noise. *International Journal of Audiology, 54*(10), 705-713.

Steel, M. M., Papsin, B. C., & Gordon, K. A. (2015). Binaural fusion and listening effort in children who use bilateral cochlear implants: A psychoacoustic and pupillometric study. *PLoS ONE, 10*(2), e0117611.

Stenfelt, S., & Rönnberg, J. (2009). The Signal-Cognition interface: Interactions between degraded auditory signals and cognitive processes. *Scandinavian Journal of Psychology, 50*(5), 385-393.

Strauß, A., Wöstmann, M., & Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Frontiers in Human Neuroscience,* 8: 350.

Strauss, D. J., & Francis, A. L. (2017). Toward a taxonomic model of attention in effortful listening. *Cognitive, Affective, & Behavioral Neuroscience*, 1-17.

Tuladhar, A. M., Huurne, N. t., Schoffelen, J. M., Maris, E., Oostenveld, R., & Jensen, O. (2007). Parieto-occipital sources account for the increase in alpha activity with working memory load. *Human Brain Mapping, 28*(8), 785-792.

Uchida, Y., Nakashima, T., Ando, F., Niino, N., & Shimokata, H. (2003). Prevalence of self-perceived auditory problems and their relation to audiometric thresholds in a middle-aged to elderly population. *Acta oto-laryngologica, 123*(5), 618-626.

Uhlmann, R. F., Larson, E. B., Rees, T. S., Koepsell, T. D., & Duckert, L. G. (1989). Relationship of hearing impairment to dementia and cognitive dysfunction in older adults. *The Journal of the American Medical Association, 261*(13), 1916-1919.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science, 283*(5401), 549-554.

Ventry, I. M., & Weinstein, B. E. (1982). The hearing handicap inventory for the elderly: a new tool. *Ear and Hearing, 3*(3), 128-134.

Ward, L. M. (2003). Synchronous neural oscillations and cognitive processes. *Trends in Cognitive Sciences, 7*(12), 553-559.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, 7: 345.

WHO. (2001). International classification of functioning, disability and health: ICF: World Health Organization.

Wilhelm B., Wilhelm H., Lüdtke H. (1999). Pupillography: Principles and applications in basic and clinical research. In Kuhlmann J., Böttcher M. (Eds.), *Pupillography: Principles, methods and applications* (pp. 1–11). München, Germany: Zuckschwerdt Verlag.

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing loss in older adulthood what it is and how it interacts with cognitive performance. *Current Directions in Psychological Science, 14*(3), 144-148.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing, 36*(4), e153-e165.

Wisniewski, M. G., Thompson, E. R., Iyer, N., Estepp, J. R., Goder-Reiser, M. N., & Sullivan, S. C. (2015). Frontal midline θ power as an index of listening effort. *Neuroreport, 26*(2), 94-99.

Wöstmann, M., Herrmann, B., Wilsch, A., & Obleser, J. (2015). Neural alpha dynamics in younger and older listeners reflect acoustic challenges and predictive benefits. *The Journal of Neuroscience, 35*(4), 1458-1467.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing, 32*(4), 498-510.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: neural correlates of pupil size as a measure of cognitive listening load. *Neuroimage, 101*, 76-86.

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*(3), 277-284.

Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., & Rönnberg, J. (2014). Cognitive processing load during listening is reduced more by decreasing voice similarity than by increasing spatial separation between target and masker speech. *Frontiers in Neuroscience*, 8: 88.

Appendix of this thesis has been removed as it may contain sensitive/confidential content