8-7-2020

Measuring listening effort using physiological, behavioral and subjective methods in normal hearing subjects: Effect of signal to noise ratio and presentation level

Lakshmi Magudilu Srishyla Kumar
James Madison University

Measuring Listening Effort Using Physiological, Behavioral and Subjective Methods in Normal Hearing Subjects: Effect of Signal to Noise Ratio and Presentation Level

Lakshmi Magudilu Srishyla Kumar

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Communication Sciences and Disorders

August 2020

FACULTY COMMITTEE:

Committee Chair: Ayasakanta Rout, Ph.D.

Committee Members/ Readers:

Rory DePaolis, Ph.D.

Yingjiu Nie, Ph.D.

To,

MY MOTHER NEELA and Dr. AYASAKANTA ROUT

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my advisor, Dr. Rout. Pursuing a doctoral degree was just a wish and Dr. Rout's constant support and guidance is what made this journey possible. You are a good teacher and mentor and above all a great human being. You have taught me how to care for people and how to be sensitive to student's needs. Thank you for all the encouragement and support.

Dr. Nie and Dr. DePaolis, it was a great experience working with you both. You have always entertained my questions and given your precious time to discuss the project findings and every good thing that happened at JMU. Thank you for being considerate about the situational needs and for helping to complete this project on time.

Dr. Meixner, Dr. Finney, Dr. Gray and Dr. Horst, you all have redefined the meaning of teaching for me. I aspire to be a teacher like you in future. Thank you for being the great teachers. Dr. O'Donoghue, Dr. Dudding, Dr. Clinard, Dr. Kuo, and Dr. Kamarunus, thank you for your timely help and constant encouragement.

I would like to extend my gratitude to Nike, Dr. Kret and Dr. Sjak-Shie for matlab codes. To all the participants for your time and cooperation.

To Dr. Rout's family, Juli didi, Mihika and Mrigank thank you for being the family here at Harrisonburg. I have really enjoyed spending time with you all, the delicious food and movie nights. The sense of security that you gave me is immense. I wish you good in fortune.

Danielle, I am happy to have met a friend like you. You are such a pleasant person to be around. Your energy is infectious. Thank you for the constant support and for being there in all good and bad times. The sweet memories that we share are the treasures of my life. You have great potential and you are a good leader. I wish good luck for all your future endeavors.

Julian, I have really enjoyed your company. You have been my savior in the times of COVID-19. It has always been a great pleasure to share my culture and cuisines with you and your family, your father and mom. You made me realize that I am a cat person. Chloe and Oscar have a great contribution in keeping my sanity intact in the difficult times. Thank you for making me a part of your family. Good luck for your future.

Ariana, Michelle, Lindsay, Heesung, Daniel it was great pleasure to have worked with you all. Wish you all good luck.

Rashmita and Sonali you have become a part of my family here at Harrisonburg in a very short duration. You both have a lot in common. Simple yet beautiful heart, hardworking nature, great humor, the list is long. I believe we have a long journey ahead to share. I wish all your dreams come true.

Amma, akka, Kala, Chikki, Saanu and Indu. The word ‘thank you’ is not sufficient to express what I feel for you. All of you have believed in me more than I did. Amma, you are the reason for what I am today. You always dreamt big for us and you fought with the world to teach us self-worth. Akka, you have taught me responsibility and virtues which I hold dear to my heart and you have always lent hand to hold on to. Kallu, you always let me go my way, but remained a strong anchor when I fly high. Chikki you have taught me

to show gratitude to people for their self-less acts. Loads of love for being the strongest women in my life. Saanu, you are my life. Dear Indu, you are a big source of confidence and inspiration for me. You have always reset my life in the difficult times and given me courage to continue and complete the work at hand. Love you a lot and wish you achieve all your goals in your life.

I would also like to extend my gratitude to AIISH and all those who held my hands and took me closer to this shore called Ph.D. The list of people who have helped me is long. Though I have not mentioned your names here, you all remain in my thoughts and prayers.

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The main objective of the study is to compare the effectiveness of pupillometry, working memory and subjective rating scale —the physiological, behavioral, and subjective measures of listening effort— at different signal to noise ratios (SNR) and presentation levels: when administered together. Eleven young normal hearing individuals with mean age of 21.7 years (*SD*=1.9 years) participated in the study. The HINT sentences were used for speech perception in noise task. The listening effort was quantified using peak pupil dilation, working memory, working memory difference, subjective rating of listening and recall effort. The rating of perceived performance, frustration level and disengagement were also obtained. Using a repeated measure design, we examined how SNR (+6 dB to -10 dB) and presentation level (50- and 65-dB SPL) affect listening effort. Tobii eye-tracker software and custom MATLAB programing were used for stimulus presentation and data analysis. SNR had significant effect on peak pupil dilation, working memory, working memory difference, and subjective rating of listening effort. Speech intelligibility had significant correlation with all of the listening effort measures except working memory difference. The listening effort measures did not correlate significantly when controlled for speech intelligibility indicating different underlying constructs. When effect sizes are compared working memory (η^2_p = 0.98) was most sensitive to SNR effect, followed by subjective rating of listening effort (η^2_p = 0.84), working memory difference (η^2_p = 0.52) and peak pupil dilation (η^2_p = 0.40). Only peak pupil dilation showed significant effect of presentation level. The physiological, behavioral and subjective measures of listening effort have different underlying constructs and the sensitivity of these measures varies in representing the effect of SNR

and presentation level. The individual data trend analysis shows different breakdown points for physiological and behavioral and subjective measures. There is a need to further explore the relationship of listening effort measures across different SNRs also how these relationship changes in persons with hearing loss.

INTRODUCTION

Listening fatigue is a common complaint presented by persons with hearing loss and they are more prone to fatigue than normal hearing individuals (Alhanbali, Dawes, Lloyd, & Munro, 2017; Kramer, Kapteyn, & Houtgast, 2006). Fatigue is broadly defined as a feeling/state of mood which results in decreased performance. It is a complex construct whose definition depends on the discipline in which it is studied (Hornsby & Kipp, 2016). Listening effort is used as an indirect measure because the underlying construct of listening fatigue is unclear. Listening effort is defined as the mental effort experienced during a listening task due to the deliberate allocation of mental (or cognitive) resources. It is hypothesized that increased listening effort causes listening fatigue or a general loss of vigor (Hornsby, 2013; Pichora-Fuller et al., 2016).

Measuring listening effort in addition to speech perception ability provides a way to quantify the effect of cognitive load on communication. Cognition is a new dimension in hearing assessment protocol. The regular hearing test battery involves pure-tone audiometry and speech perception measures, which quantify the effect of hearing loss on tone detection and speech perception in quiet/noise. However, the process of communication is more complex. The cognitive resource allocation underlying the process of communication influences the ease of communication. According to Ease of Language Understanding (ELU) theory, adverse listening conditions lead to increased need for spending cognitive resources (Rönnberg et al., 2013; Rönnberg, Rudner, Foo, & Lunner, 2008) resulting in increased listening effort. Measuring listening effort can shed light on mental effort that a person experiences during adverse listening conditions and

resulting listening fatigue. Also, it may help explain inter-subject variability in speech recognition scores.

The increased listening effort and fatigue are shown to have adverse effects on quality of social and work life of persons with hearing loss leading to social isolation (Gosselin & Gagné, 2011; Kramer et al., 2006; Pichora-Fuller, Mick, & Reed, 2015). In the United States, one in eight or 30 million people who are 12 years or older have hearing loss in both ears (Lin, Niparko, & Ferrucci, 2011). The prevalence of hearing loss increases with age- 25% of those aged 65 to 74 years and 50% of those aged 75 years or older have disabling hearing loss (35 dB or greater in better ear) (NIDCD, 2016). The elderly persons with hearing loss are more susceptible to the effects of listening effort due to age related cognitive decline (Gosselin & Gagné, 2011a; Gosselin & Gagné, 2011b; Meister et al., 2013). Listening effort is a sensitive measure to investigate the benefit of hearing aids and cochlear implants compared to speech intelligibility measures (Johnson, Xu, & Cox, 2016; Pals, Sarampalis, & Baskent, 2013; Sarampalis, Kalluri, Edwards, & Hafter, 2009; Winn, 2016). Hence, measuring listening effort has significant clinical relevance.

The listening effort tools can be grouped into three classes: (1) the physiological measures, (2) the behavioral methods, and (3) the subjective measures. The physiological methods measure the arousal response of the autonomic nervous system (like pupil dilation, heart rate, and skin resistance) or electrophysiological responses of brain in response to cognitive load. The behavioral methods examine the changes in task performance during listening task with and without cognitive load. The change in performance is believed to be the result of increased mental effort. Subjective rating of

effort is another listening effort measurement method. These tools assess the level of effort involved in a specific listening condition or the general listening effort in day-to-day life.

Pupillometry is a physiological measure of pupil dilation mediated by activity in locus coeruleus (a noradrenergic system hub). The pupil dilation is modulated by changes in attention, stress and memory load (McGarrigle et al., 2014; for more information refer Laeng, Sirois, & Gredebäck, 2012). Pupillometry is shown to be sensitive to listening task difficulties such as changes in signal to noise ratio, spectral distortion, and contextual load (Pals et al., 2013; Winn, 2016; Zekveld & Kramer, 2014). Hence, it is considered one of the more reliable tools among all the physiological measures.

The listening span is a behavioral test which combines recall task with speech recognition task. According to ELU theory, in adverse listening condition the working memory is recruited to process degraded speech (Shehorn, Marrone, & Muller, 2018). As the task difficulty increases, the speech recognition task recruits most of the resources leaving little resources to store the information, thus reducing the recall ability (Pichora-Fuller, Schneider, & Daneman, 1995; Sarampalis et al., 2009). Listening span is sensitive to signal-to-noise ratio changes, contextual information, aging, absence or presence of noise reduction strategy (Johnson, Xu, Cox, & Pendergrafta, 2015; Pichora-Fuller et al., 1995; Sarampalis et al., 2009). Listening span test is easy to incorporate with the speech recognition tests currently in use and has ecological validity as testing mimics the everyday communication situation.

The subjective rating methods show increase in perceived effort with increase in task difficulty. The subjective rating scale is sensitive to the effect of task load such as speech perception in various levels of noise, quite versus noise, aided condition versus unaided condition and hearing aid processing strategies settings (Bentler, Wu, Kettel, & Hurtig, 2008; Brännström, Karlsson, Waechter, & Kastberg, 2018; Pals et al., 2013; Rudner, Lunner, Behrens, Thorén, & Rönnberg, 2012; Strand, Brown, Merchant, Brown, & Smith, 2018). The subjective rating is also shown to be sensitive to internal factors like presence of hearing loss (Desjardins & Doherty, 2014; Humes, Christensen, Bess, & Hedley-williams, 1997). Among all the measures of listening effort subjective rating is the easiest method to administer and assess the perceived effort in clinical settings. It is also cost effective and has good face validity as it examines the person's perception of a situation. However, there is wide variation in the scales used in researches. The currently available rating scales are either borrowed from other disciplines or sub-tests adapted from existing tests and are not specifically developed for the purpose of measuring listening effort (Alhanbali et al., 2017; Hughes, Rapport, Boisvert, McMahon, & Hutchings, 2017). There are no standardized subjective rating scale available for clinical use (Hughes, Hutchings, Rapport, McMahon, & Boisvert, 2018). Moreover, the construct behind the questions used to measure listening effort is not clear. Nevertheless, subjective ratings are time efficient tools that can be easily included in the assessment battery.

Despite the evidence of increased listening fatigue and effort in persons with hearing loss, and potential application of using these measures to assess rehabilitation in clinics, several questions need answers. It is not clear as to which measure is more sensitive to task difficulty such as changing signal to noise ratio and presentation level.

Furthermore, it is not clear if different tools provide the same information on task difficulty. Study by Strand et al. (2018) examined seven tools to seek evidence to ELU theory including pupillometry, listening span and subjective rating scale. All the three measures were sensitive to task difficulty (signal to noise ratio). The convergent validity analysis showed a significant weak positive correlation between listening Span and subjective rating measure and a significant weak negative correlation between pupillometry and listening Span. There was no significant correlation between pupillometry and subjective rating. The different direction and small magnitude of relationships were considered as evidence of different underlying construct.

Framework for Understanding Effortful Listening (FUEL) is a working model developed to explain the process of listening effort (Pichora-Fuller et al., 2016). According to this conceptual framework, the physiological tests capture the involuntary arousal response in the autonomic nervous system in reaction to difficult listening situations and moment to moment variation in cognitive load during a task. The effect of resource allocation to store and process auditory information during adverse listening conditions is captured using working memory tests (Pichora-Fuller, 2010). The subjective measures give information on the person's experience of effortful listening after going through the task. Though Strand et al. (2018) showed different constructs of the tools, the study used only two SNR conditions. As SNR and listening effort measures are shown to have non-linear relationship (Ohlenforst et al., 2017; Zekveld & Kramer, 2014), just two conditions may not be sufficient to establish how pupillometry, working memory and subjective rating scales are differently sensitive to task difficulty (signal to noise ratio).

The current project aims to see how listening effort measured using pupillometry (a physiological measure), working memory test (behavioral measure), and subjective rating (subjective measure) changes across six different signal-noise ratio (SNR) conditions and presentation levels. Speech perception in noise task is used to manipulate task difficulty as it is the most challenging situation for persons with hearing loss and noise is the most common factor that affects speech clarity. An extended SNR range is used to trace how three different measures of listening effort change as a function of task difficulty. Furthermore, concurrent measurement using three tools within each condition should help to control extraneous variable like state of mind which may influence the results of pupillometry. The study plans to examine the effect presentation level on listening effort because audibility of stimulus is another major factor which decides the success of rehabilitation options such as hearing aids and cochlear implants. Previously very few studies have examined the effect of presentation level on listening effort measures. A study by Liao, Kidani, Yoneya, Kashino, and Furukawa, (2016) examined the effect of presentation level of tones, noise-bursts on pupil response and observed louder signals to be associated with larger pupil responses. In contrast, a study by Zekveld et al. (2010) the baseline level did not vary significantly when pupil responses were measured at different noise levels while keeping the sentence level constant. In addition to equivocal results with respect to presentation level effect, there are no studies which examine the effect of presentation level on peak pupil dilation change in the time window where a person is listening to speech in noise. In the present study, the effect of presentation level will be examined on pupil dilation, working memory, and subjective

measure. Examining the effect of presentation level may help to hypothesize the effect of different suprathreshold gain in persons with hearing impairment.

The second aim of the study is to compare the sensitivity of three listening effort measures to examine which measure is more sensitive to changes in SNR and presentation level. Having the information on test efficacy is critical for the clinical adaptation of listening effort measures. There is an abundance of measures that quantify listening effort in the literature. The three major classes of measures are physiological measures, behavioral measures and subjective rating measures.

Various studies have examined the comparative sensitivity of listening effort measures and have shown equivocal results (Alhanbali, Dawes, Millman, & Munro, 2019; Johnson et al., 2015; Seeman & Sims, 2015). Alhanbali and colleagues (2019) simultaneously measured pupil size, electroencephalographic alpha power, skin conductance, and self-reported measure of effort in 116 participants with normal to severe hearing loss. The testing was conducted at SNR corresponding to 71% performance on digit recall task and results showed pupillometry to explain higher percent of variance compared to alpha power changes and subjective rating. Study by Johnson et al, (2015) showed subjective rating scale to be more sensitive in reflecting changes in SNR than listening span test in normal hearing individuals (N=30). Seeman and Sims (2015) compared physiological measures (skin conductance, hear rate, and heart-rate variability), dual-task measure and subjective ratings at two SNRs (+5 and +15 dB) in normal hearing individuals and found subjective rating compared with physiological or dual-task measure to be more sensitive.

It is difficult to select tools for clinical use because: the studies examine tools efficiency at a small range of task difficulty (Alhanbali et al., 2019; Seeman & Sims, 2015); the studies compare tools with selective underlying constructs (Johnson et al., 2015). Comparing the physiological, behavioral and subjective measures is important as they have different strengths and weaknesses. The physiological measures provide temporal precision and effort change across time, but these measures require dedicated equipment. The behavioral measures represent real life experience, but internal factors like ability to perform multiple tasks, baseline working memory capacity may affect the results. The subjective measures give face validity as it measures the experience of persons but may get influenced by the subjective bias and misperception of the questions.

The present study aims to examine the efficiency of pupillometry, working memory, and subjective rating scales in demonstrating the effect of signal to noise ratio and presentation level. The pupillometry was selected as it has shown consistent pattern of results for the effect of SNR. The subjective rating scale was selected as it is most easy to administer, cost effective and time efficient measure. The working memory test was selected as it is an easy behavioral measure to incorporate along with already existing sentence perception task in the clinics. Also, these three measures were selected as they represent different classes of listening effort measurement methods.

METHODS

Study design

The current study used a two-way repeated measures experimental design. The participants for the study were selected using a non-random convenient sampling method. A power analysis conducted using G*Power 3.1.9.2 software (Faul, Erdfelder, Lang, & Buchner, 2007) with medium effect size (*cohen's f* = 0.25) at 0.05 alpha level indicated that a minimum of 14 participants is required for the study to achieve 0.8 power. The independent variables were signal to noise ratio (six conditions) and presentation level (two levels). The dependent variables used to measure listening effort were pupillometry, working memory and a subjective rating scale of listening effort and recall effort.

In pupillometry, the parameter of interest was peak pupil dilation. In working memory, the number of words correctly recalled was used as a measure of listening effort. In addition, the working memory difference or memory cost was measured as the difference between number of correctly recalled words out of number of possible answers. The two questions which estimated the effort in listening and recall tasks were considered dependent variables. The questions that measured frustration level, disengagement and performance were used as co-variates. The SNRs for each participant were counterbalanced using Latin square method to minimize order effect. For the first presentation level, the order used was 1, 2, n, 3, 4, 5, (n-1) where n is the highest number of the condition (six in the present study). For the second presentation level, this order was reversed to get a new sequence. For the successive participants , the sequence was decided by adding one to each condition in the previous subject's sequence and by

replacing the highest order condition with one. For the second presentation level, the conditions were reversed to create a new sequence.

Participants

A total of 14 participants enrolled in the study and out of the 14, eleven participants completed the testing. The participants were native speakers of American English and had pure tone thresholds within 20 dB at octaves within 250 to 8000 Hz. range Hearing thresholds were obtained with a GSI Audiostar Pro audiometer using THD-49 supra aural headphones calibrated in accordance with ANSI S3.6-1996. Normal middle ear function was evaluated by confirming a type 'A' tympanogram using an Interacoustics instrument. The mean age was 21.7 years (SD = 1.9 years) and 10 participants were female, and one was male participant. The participants did not have any past history of eye injury or congenital eye problems, attention disorder, epilepsy, recent history of middle ear problem or self-reported difficulty of speech perception in noise or were under any medications at the time of testing. Two participants who completed the study had corrected vision. The study protocol was reviewed and approved by Internal Review Board of James Madison University. The participants were paid $20 compensation for their participation.

Stimuli

Speech Perception and Working Memory

Sentences from the Hearing in Noise Test (HINT) (Nilsson et al., 1994) with speech shaped noise were used for speech perception task in noise. The sentences were presented at six signal to noise ratios (SNR) ranging from +6 dB SNR to -10 dB SNR.

The SNRs were 3 dB apart except for -10 dB which was 4 dB lesser than -6 dB condition. The different SNR conditions were generated using MATLAB code (Nike, 2017). The SNR was calculated based on RMS amplitude of the signal. The RMS amplitude of speech was calculated with the natural pauses inside the sentence intact. Before mixing the sentences and noise, the RMS level of sentences were kept constant and then the required noise level was calculated based on the SNR ($RMS_{Noise} = RMS_{Speech} - SNR$). The sentences were then added to the noise to create different SNR conditions. Three seconds of noise was inserted before and after the sentence to monitor the trajectory of pupillometry. The level of the noise before and after the sentence increased with reduction in SNR. The level of speech mixed with noise was maintained constant across SNRs.

The HINT sentences were used to measure the working memory or listening span. The last word recall task was used to measure the working memory of subjects. The HINT sentences were arranged in blocks of five sentences (four blocks in each SNR condition). In each condition there were a total of 20 sentences. Different sentence lists were used for two different presentation levels. The sentence lists were counterbalanced between presentation levels to avoid any systematic effect of the lists.

Tone detection test

A tone detection test was included at SNRs ranging from +6 dB to -10 dB to separate the effect of linguistic context present in the HINT sentences. It was hypothesized that if the presentation level effect is due to just loudness both speech recognition task and tone detection task would show the effect of presentation level;

whereas, if it is due to speech understanding or task engagement reasons (linguistic context), tone detection task would not show the effect of presentation level. A 1000 Hz pure tone and 1/3rd octave narrow band noise were generated using an audiometer and were recorded using Sound Forge 9 software (Sony Digital Audio) to create signals with different SNRs. The different SNR conditions were generated using MATLAB code (Nike, 2017). The procedure to add tone and noise and arrangement of stimulus were all similar to speech stimulus preparation methods. The stimuli had two second baseline (silence) before the onset of noise. The two second tone was embedded in the center of an eight second noise.

Subjective questionnaire

The subjective questionnaire to measure perceived effort of participants was adopted from the NASA-TLX questionnaire (Hart, 2006; Hart & Staveland, 1988) and Effort Assessment Scale (Alhanbali, Dawes, Lloyd, and Munro, 2018) . The short questionnaire included five questions, where two questions measured effort due to listening to speech in noise and remembering/recalling words. The other three questions measured the performance, frustration and disengagement from the task. The questions were rated on a ten-point rating scale where a rating of 1 indicated low effort, frustration, disengagement and high performance, and a rating of 10 indicated high effort, frustration, disengagement and low performance.

Pupillometry

For pupillometry, the HINT sentences were converted into videos with gray background and then were arranged inside the Tobii Studio software (Tobii Pro AB,

Stockholm, Sweden). The videos had two second silence before the beginning of the stimulus to serve as the baseline. Five seconds of interstimulus interval was provided to return the pupil size back to stable baseline. The participants were provided with a five seconds gap to repeat the sentence and a maximum of fifteen seconds to recall the words (Appendix. 2).

Instruments

The testing was conducted in a sound attenuated room. The Tobii T60 XL screen based eye tracker (Tobii Pro AB, Stockholm, Sweden) was used to measure the pupil diameter. A personal computer with Tobii Studio placed outside the sound booth was used to control the presentation of the stimulus and collecting pupillometry data. The eye tracker had a sampling rate of 60Hz and used infra-red rays to measure the pupil dilation. The participants were seated approximately 65 cm away from the eye-tracker screen. The participants were provided with a chin rest to stabilize the head position. This helped to keep the distance between screen and head of the participant constant across conditions.

The sentences were routed through a GSI Audiostar pro audiometer (Grason-Stadler, Eden Prairie, MN) to two loudspeakers placed ear level at 45^0 angles inside the sound attenuated booth and were presented at 50 dB and 65 dB SPL. The presentation levels were calibrated using a Quest SoundPro class I sound level meter (TSI Inc., Shoreview, MN) before the testing commenced for each participant to match the target presentation levels at head level. The brightness of the room was kept constant throughout the testing and across participants. Before each session, the researcher made sure that the luminance was at its maximum using a dimmer switch.

Procedure

The study participants were recruited through flyers posted at different locations around James Madison University campus. Once the participant showed interest in participating in the study, a questionnaire was sent through email containing consent form and a questionnaire related to inclusion and exclusion criteria. The questionnaire also included questions on recent ear infection, self-reported problem of speech perception in noise, and musical training. The participants were excluded from the study if they had any past history of eye injury or congenital eye problems, attention disorder, epilepsy, recent history of middle ear problem or self-reported difficulty of speech perception in noise or were under any medications at the time of testing. If the respondent met all inclusion criteria, they were contacted again to inform their selection into the study and to schedule an appointment for testing. The testing was conducted in two sessions. In the first session the participant underwent a hearing screening, a practice session to get familiarized with the task and speech recognition and tone detection testing at one presentation level. The first session took approximately two hours fifteen minutes. Hearing screening included pure-tone audiometry and immittance screening. The practice condition was done at +15 dB SNR with ten sentences to familiarize the procedure to participant. In the second session, testing was conducted at the second presentation level at six SNRs. The second session lasted approximately two hours. The presentation levels were counterbalanced to minimize any order effect.

Speech perception, pupillometry and working memory

The participants underwent a total of twelve conditions of speech perception in noise. All the SNR conditions were presented at two presentation levels that is 50 dB and 65 dB SPL. Each condition had twenty sentences each. The participants were instructed to repeat the sentence they heard. A visual prompt was displayed on the eye tracker screen at the end of every sentence to repeat the sentence. After every five sentences, the participants were cued to recall the last words of each of the sentences. There were five seconds time to repeat the sentence and a maximum of fifteen seconds to recall the words. The participants were encouraged to guess responses when needed. The experimenter switched the stimulus after recall response at the end of every five-sentence block. The participant's pupil dilation was monitored throughout the speech perception and recall tasks to measure the changes in pupil dilation corresponding the speech perception task and recall task. The experimenter and another trained audiologist scored the sentence recognition and recall responses during the testing. The complete testing session was video recorded for later offline speech perception and working memory scoring.

Subjective rating

The participants rated effort after each condition using the listening effort questionnaire. The experimenter checked with the participants if they needed break after each condition and a five-minute break was provided whenever desired. Each condition took approximately eight minutes to complete.

Data Analysis

Pupillometry analysis

Original pupillometry data analysis MATLAB codes by Kret and Sjak-Shie (2019) were modified to analyze the pupil data in the current study. The preprocessing of pupil data included three steps. In the initial stage the data was filtered using a range filter, a speed filter and a deviation filter to remove eye blink artifacts and isolated islands of data. The range filter removed any pupil data which was outside 1.5 to 9 mm range. The speed filter was used to remove eye blinks which resulted in a sudden change in pupil diameter. The speed was calculated as the ratio of unit change in pupil diameter to unit change in time (Equation 1). A median absolute deviation (MAD) method was used to remove the outliers (Equation 2). The threshold for outlier removal was calculated using the following formula (Equation 3) (Kret & Sjak-Shie, 2019). The median (d') was the median of the speed calculated for the adjacent pupil data points in both directions. The n for threshold calculation were selected after visual inspection of the data post filtering. The deviation filter used the same MAD method for removing the saccadic artifact and spurious islands of data between gaps.

$$d'^{[i]} = \max\left(\left|\frac{d[i]-d[i-1]}{t[i]-t[i-1]}\right|, \left|\frac{d[i+1]-d[i]}{t[i+1]-t[i]}\right|\right) \quad \text{(Equation 1)}$$

$$MAD = median(|MAD = median(|d' - median\ (d')|) \quad \text{(Equation 2)}$$

$$Threshold = median\ (d') + n.MAD \quad \text{(Equation 3)}$$

After processing the pupil data, the valid samples were retrieved, and the percentage of data remained after processing were calculated. If a trial had less than 30%

of samples left, it was removed from further analysis. The peak pupil dilation was selected within each trial of speech perception (onset to offset of the stimulus) were calculated to be used for future statistical analysis.

Speech perception and Working memory

Speech perception scores were measured in two metrics -1) the number of sentences correct out of twenty and 2) the proportion of sentences correct. Two Audiologists scored the responses independently during the testing and came together to compare the responses. If there was any discrepancy, they reanalyzed the video recorded response to arrive at a consensus. The sentences were scored correct only when all the words in a sentence were perceived correctly.

Working memory was calculated as the proportion of words correctly recalled out of twenty words for each condition. When a sentence was misperceived, the recall score was still awarded if the participant repeated a complete sentence with length matching ±2 functional words of the original sentence and recalled the last word as they perceived. Working memory difference score was measured as the difference between number of words correctly recalled and the number of possible answers. A possible answer was defined as a 50% correctly identified grammatically complete sentence with length within ±2 words of original sentence.

Reliability of the measures

To evaluate the reliability of pupillometry, working memory and subjective rating measures, the original testing protocol included retesting of 20% of the participants. The

stimulus and testing procedures were kept same. Due to the COVID-19 pandemic, data collection was suspended. No reliability data were collected at the time of this writing.

RESULTS

Effectiveness of individual listening effort measures, trend analysis and comparison of listening effort measures using graphical methods and correlation analysis is presented in this section. The listening effort data was analyzed for pupillometry, working memory and subjective ratings separately, followed by trend analysis and comparison and correlation analysis. The effectiveness of measures in depicting the effect of SNR and presentation level is evaluated using two-way repeated measures ANOVA. For pupillometry data, due to small sample size both group data and individual data are analyzed and presented.

An *a priori* power analysis indicated 14 participants are required to have 0.80 power with medium effect size at alpha level of 0.05. The data was collected from 11 participants due to COVID-19 restrictions on research activities. Out of the eleven participants, five participants had complete data in all conditions from pupillometry. Four of the eleven participants had missing data in some of the conditions and two participants did not have any valid data for analysis. There were total six two-way repeated measures analysis conducted on six dependent variables. To control for familywise error, the p value was adjusted by dividing 0.05 by 6. The new alpha level used was 0.008. The pairwise comparisons exploring main effect of SNR and presentation level and interaction were evaluated at p=0.05 with Bonferroni correction.

I. Pupillometry

The pupil response represents pupil diameter change from baseline while listening to speech in noise. Greater change indicates more listening effort. As shown in Figure 1,

listening effort increased gradually for both 65 dB and 50 dB presentation levels. The listening effort was highest at -6 dB SNR for both presentation levels and there was a drop in effort at -10 dB SNR at both presentation levels. The pupil dilation was also higher for 65 dB compared to the 50 dB presentation level.

Table 1 Mean, SD and range of peak pupil dilation change across SNRs and presentation levels

Descriptive Statistics

	50 dB SPL						65 dB SPL					
	6	**3**	**0**	**-3**	**-6**	**-10**	**6**	**3**	**0**	**-3**	**-6**	**-10**
Valid	8	7	8	7	8	8	9	9	8	8	9	9
Missing	3	4	3	4	3	3	2	2	3	3	2	2
Mean	0.41	0.40	0.42	0.45	0.49	0.49	0.43	0.43	0.46	0.47	0.58	0.54
Std. Deviation	0.07	0.05	0.06	0.10	0.16	0.14	0.07	0.10	0.11	0.10	0.17	0.11
Minimum	0.19	0.32	0.35	0.36	0.29	0.32	0.19	0.24	0.25	0.26	0.30	0.25
Maximum	0.48	0.46	0.51	0.62	0.63	0.63	0.53	0.53	0.58	0.59	0.78	0.70

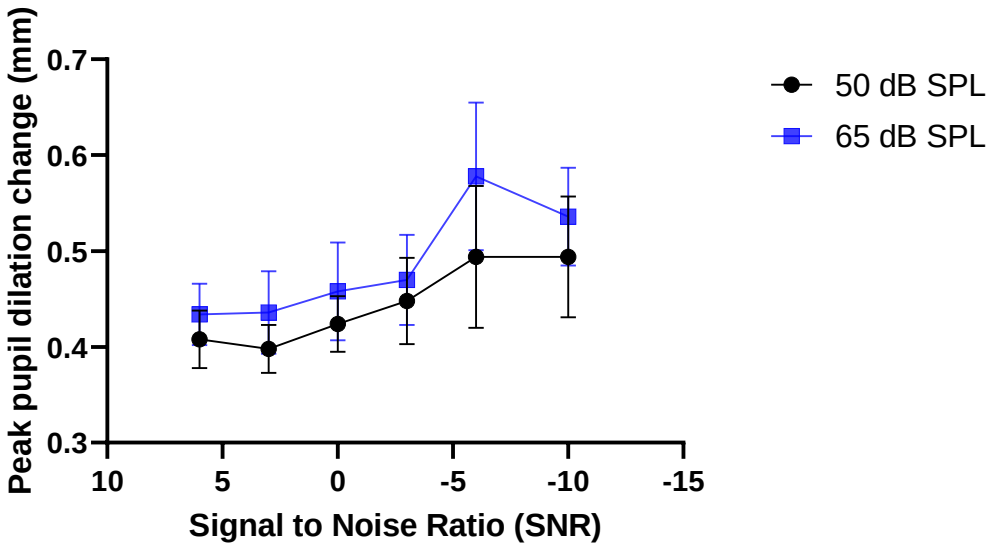

Figure 1 Peak pupil dilation change (in millimeters) across different SNRs and presentation levels. The error bars represent ±1 SE.

Individual pupillometry data analysis

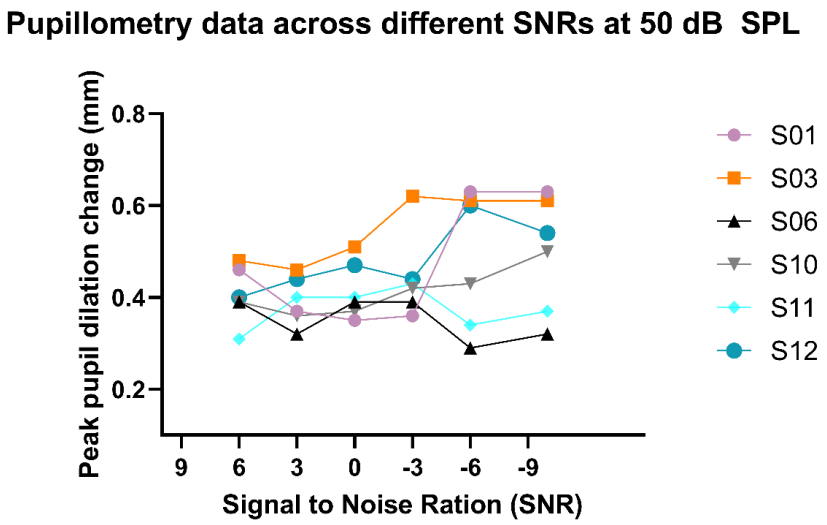

Figure 2 Peak pupil dilation change (in millimeters) across different SNRs at 50 dB SPL (Individual data)

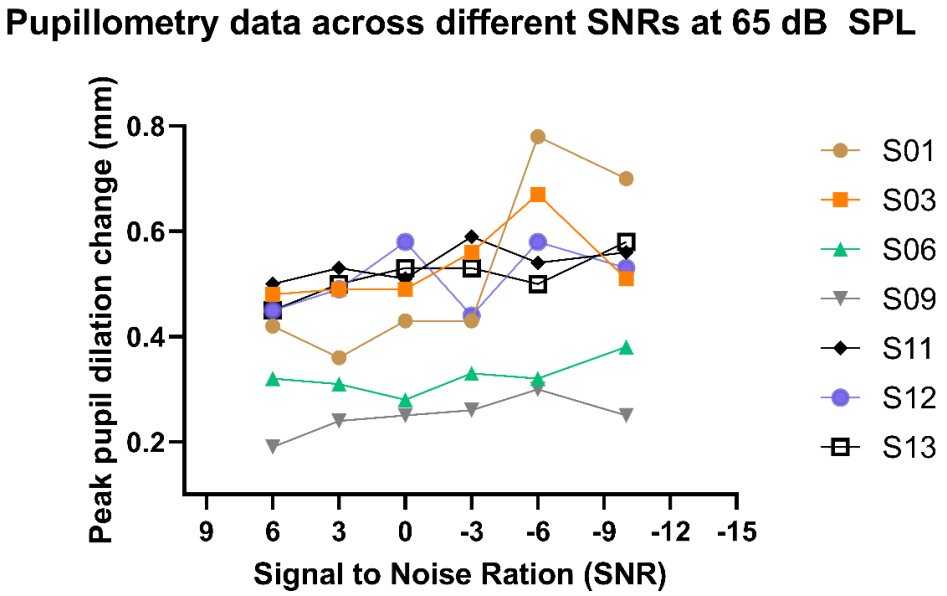

Figure 3 Peak pupil dilation change (millimeters) across different SNRs at 65 dB SPL (Individual data)

The trend in individual data set was analyzed using graphical methods. The peak pupil dilation change across different SNRs and presentation levels are shown in Figure 2 and 3. There was high variability in the magnitude and trend of dilation change across participants. The mean effort change ranged between 0.2 to 0.6 mm. For some participants, the listening effort increased with increase in speech understanding difficulty up to or at a certain point (-3 dB or -6 dB SNR) and then dropped off or saturated at more negative SNRs. For example, S01 showed an increase in effort at -6 dB SNR and -10 dB SNR at both 50- and 65-dB SPL. S03 shows an increase at -3 dB SNR, 50 dB SPL and at -6 dB SNR, 65 dB SPL. For majority of the participants there was not enough variation in the listening effort across different SNRs. Another participant's data (S12) showed irregular pattern in listening effort at 65 dB.

All five participants with complete data showed higher effort at 65 dB SPL. The magnitude of difference in effort ranged between 0.09 to 0.98 mm. The effect of presentation level was not affected by order of presentation.

Effectiveness of pupillometry: SNR and presentation level effect on listening effort

A two-way repeated measures ANOVA was administered to evaluate the effectiveness of pupillometry in examining effects of the SNR and presentation level on listening effort. The data distribution was assessed using histogram, skewness, kurtosis, and box plots for normality. The distribution at group level showed non-normal distribution. The box plots showed few outliers. The two-way repeated measures ANOVA is run ignoring non-normality as ANOVA is robust for the violation of normality assumption (Tabachnick & Fidell, 2013). The outliers were not removed as the

sample size is small and removal of data in one condition removes the entire data set reducing the power during the repeated measures ANOVA. The test was administered on complete data sets obtained from five participants. The results showed no significant interaction (p=0.60) and main effect of SNR (p=0.07) and presentation level (p=0.29) on listening effort The observed power ranged between 0.16 to 0.28 for main effects and interaction (at p = 0.05).

The two-way repeated measures ANOVA was re-administered on data by replacing the missing data with group mean (Tabachnick & Fidell, 2013). The assumption of Sphericity was tested using Mauchly's test. The results showed data to violate the assumption of Sphericity ($\chi^2(14) = 39.29$, $p < 0.0001$). Hence, the Greenhouse-Geisser correction was used to interpret repeated measures ANOVA results. The results showed a significant interaction between SNR and presentation level ($F(3.9, 38.98) = 4.63$, p=0.004, $\eta_p^2 = 0.316$). The observed power was 0.91 with eleven participants (at p = 0.05). The results indicated that the pattern of listening effort change across SNRs is different for 50- and 65-dB SPL presentation levels. The main effect of SNR was also significant ($F(1.76, 17.64) = 6.11$, p<0.0001, η_p^2 =0.38). The main effect of presentation level was not significant ($F(1,10) = 3.65$, p=0.08, $\eta_p^2 = 0.27$).

To explore the interaction, post-hoc analysis was conducted using Bonferroni correction (Table 2). The results showed significant difference in listening effort between 50- and 65-dB presentation levels at -6 dB SNR. The listening effort was higher for 65 dB presentation level compared to 50 dB (Figure 4). The presentation level was not significant at other SNRs (p>0.008).

Table 2 Pairwise comparison of pupil dilation change (listening effort) between presentation levels across SNRs.

SNR	Mean Difference (50 dB-65 dB)	Std. Error	Sig.[b]	95% Confidence Interval for Difference	
				Lower Bound	Upper Bound
6 dB	-.039*	.023	.122	-.091	.013
3 dB	-.019*	.022	.417	-.067	.030
0 dB	-.040*	.029	.195	-.104	.024
-3 dB	-.001	.025	.972	-.056	.054
-6 dB	-.095*	.028	.007	-.158	-.032
-10 dB	-.041*	.026	.150	-.099	.017

*indicate significance at 0.05.

II. Working memory

The working memory was measured using last word recall task. The working memory was quantified in two ways: the number of words recalled correctly per condition (out of 20 words)- working memory; and difference between the number of words recalled correctly out of number of possible answers- working memory difference. A possible answer was defined as a 50% correctly identified grammatically complete sentence with length within ±2 words of original sentence. The second variable was calculated to avoid the influence of audibility. The mean working memory and working memory difference are shown in Figures 4 and 5, respectively. The higher working memory scores indicate lower listening effort (Figure 4). The working memory decreased as the SNR reduced indicating increase in listening effort. The working memory difference is the number of words missed by the participant, hence, higher the number, higher is the listening effort. From Figure 5 we can notice that the listening effort increases up to -4 dB SNR and reduces at -6- and -10-dB SNR.

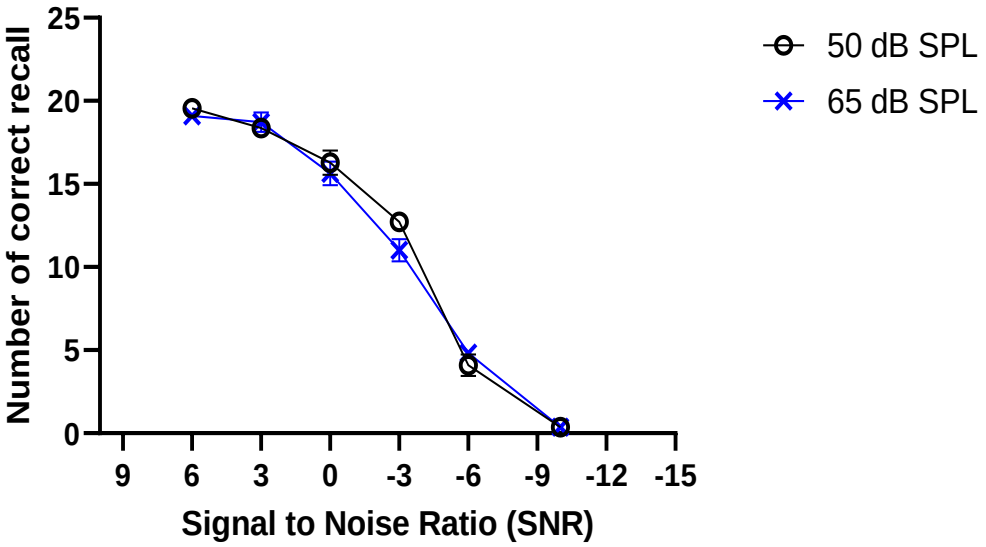

Figure 4 Working memory (out of maximum 20) across SNRs and presentation levels. The error bars represent ±1 standard error.

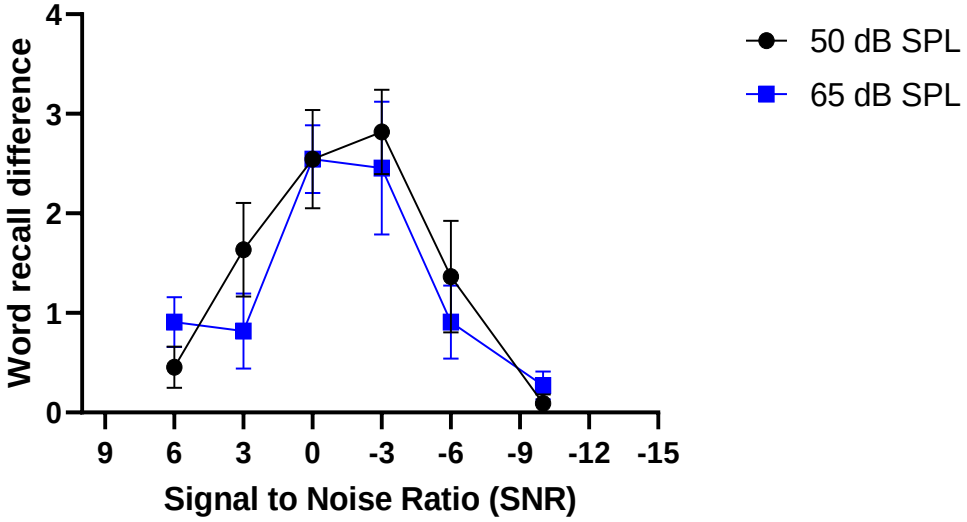

Figure 5 Working memory difference across SNRs and presentation levels. The words incorrect recall= possible correct recall – words correct recall. The error bars represent ±1 standard error.

Two-way repeated measures ANOVA was conducted to evaluate the effectiveness of working memory and working memory difference in representing the effect of SNR and presentation level (N=11). The normality assumption was assessed using histogram, skewness, kurtosis, and box plots. The distribution at group level showed non-normal distribution. The box plots showed few outliers. The two-way repeated measures ANOVA is run ignoring non-normality as ANOVA is robust for the violation of normality assumption (Tabachnick & Fidell, 2013) and as groups have equal N. The outliers were not removed as the sample size is small and to maintain power. Significant Mauchly's test of sphericity for working memory difference showed violation of sphericity assumption ($p<0.05$). Hence, Greenhouse-Geiser correction was considered during interpretation of results.

The interaction and main effect of presentation level were not significant for both working memory and working memory difference ($p>0.008$). The listening effort significantly changed with change in SNR for both working memory ($F(3.33, 32.29)=433.42$, $p<0.0001$, $\eta_p^2=0.98$) and for working memory difference ($F(2.99, 29.86)=11.02$, $p<0.0001$, $\eta_p^2=0.524$). A pairwise comparison was done using Bonferroni correction. The results are provided for working memory and working memory difference in Table 3 and 4, respectively. The listening effort differed significantly between all SNR conditions for working memory ($p<0.05$) except 6- and 3-dB SNR. Listening effort was significantly higher for 0- and -3-dB SNR (medium difficulty in speech perception) compared to 6 dB and -10 dB SNR (easiest and most difficult speech perception conditions) for working memory difference ($p<0.05$).

Table 3 Pairwise comparison of working memory across SNRs

Post Hoc Comparisons – Working memory

		Mean Difference	SE	t	Cohen's d	p_{Bonf}
6	3	0.773	0.493	1.569	0.473	1.000
	0	3.364*	0.688	4.890	1.474	0.009
	-3	7.455*	0.533	13.991	4.218	< .001
	-6	14.864*	0.472	31.466	9.487	< .001
	-10	18.955*	0.184	102.971	31.047	< .001
3	0	2.591*	0.563	4.599	1.387	0.015
	-3	6.682*	0.581	11.500	3.467	< .001
	-6	14.091*	0.583	24.163	7.285	< .001
	-10	18.182*	0.433	41.978	12.657	< .001
0	-3	4.091*	0.551	7.423	2.238	< .001
	-6	11.500*	0.647	17.783	5.362	< .001
	-10	15.591*	0.639	24.401	7.357	< .001
-3	-6	7.409*	0.534	13.865	4.181	< .001
	-10	11.500*	0.416	27.671	8.343	< .001
-6	-10	4.091*	0.425	9.616	2.899	< .001

Note. Cohen's d does not correct for multiple comparisons.
*Significant at p=0.05

Table 4 Pairwise comparison of working memory difference across SNRs

Post Hoc Comparisons – Working memory difference

		Mean Difference	SE	t	Cohen's d	p_{Bonf}
6	3	-0.545	0.413	-1.322	-0.399	1.000
	0	-1.864*	0.405	-4.601	-1.387	0.015
	-3	-1.955*	0.434	-4.503	-1.358	0.017
	-6	-0.455	0.423	-1.073	-0.324	1.000
	-10	0.500	0.165	3.028	0.913	0.191
3	0	-1.318	0.423	-3.120	-0.941	0.163
	-3	-1.409	0.571	-2.466	-0.744	0.500
	-6	0.091	0.555	0.164	0.049	1.000
	-10	1.045	0.297	3.516	1.060	0.084
0	-3	-0.091	0.436	-0.209	-0.063	1.000
	-6	1.409	0.567	2.484	0.749	0.485
	-10	2.364*	0.331	7.143	2.154	< .001
-3	-6	1.500	0.393	3.816	1.150	0.051
	-10	2.455*	0.378	6.491	1.957	0.001
-6	-10	0.955	0.372	2.566	0.774	0.422

Post Hoc Comparisons – Working memory difference

	Mean Difference	SE	t	Cohen's d	p Bonf

Note. Cohen's d does not correct for multiple comparisons.
*Significant at p=0.05

III. Subjective rating of listening effort

The subjective rating of listening effort was measured using two questions. The participants were asked to rate how difficult it is to follow and understand sentences. The results for this question was termed 'listening effort'. The participants were also asked to rate how difficult it is to remember and recall the words and the variable was termed 'recall effort'. Figures 6 and 7 show the mean listening and recall effort across different SNRs and presentation levels, respectively. The self-reported listening effort increased as the SNR worsened, for both presentation levels and the listening effort was higher by 0.72 units and 1 unit at +3 dB and -3 dB for 65 dB presentation level compared to 50 dB presentation level. The recall effort increased gradually for 50 dB presentation level as the SNR deceases. For 65 dB presentation level the effort was less for positive SNRs (+3- and +6-dB SNR) and at -10 dB SNR compared to 0, -3, and -6 dB SNR. The trend is similar to working memory difference (see figure 5) where working memory difference was the least at +6 dB and -10 dB SNRs. For recall effort the variability was high at +6- and -10-dB conditions.

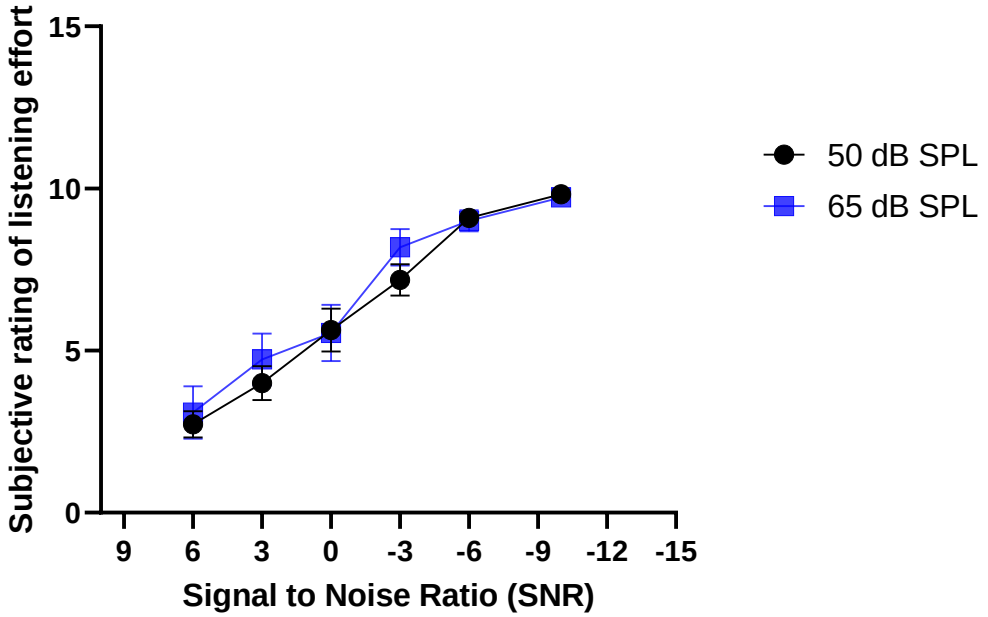

Figure 6 Subjective rating of listening effort across SNRs and presentation levels. The error bars represent ±1 SE.

Two-way repeated measures ANOVA were administered to evaluate the effect of SNR and presentation level on self-reported listening effort. Interactions and main effect of presentation level were not significant for both listening and recall effort ($p<0.008$). The main effect of SNR was significant only for listening effort ($F(2.74, 27.45) = 53.58$, $p<0.0001$, $\eta_p^2 = 0.84$). The pairwise comparisons with Bonferroni correction showed significant increase in listening effort with reduction in SNR except the 3 dB to 0 dB and -6- and -10-dB pairs. The results are as shown in Table. 5.

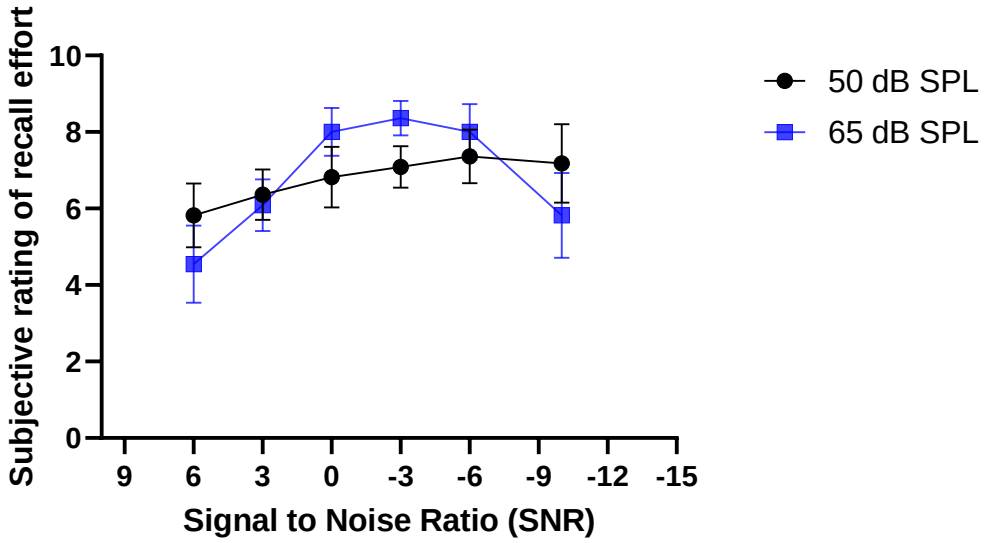

Figure 7 Subjective rating of recall effort across SNRs and presentation levels. The error bars represent ±1 SE.

Table 5 Pairwise comparison of listening effort across SNRs

Post Hoc Comparisons – SNR main effect for Listening effort

		Mean Difference	SE	t	Cohen's d	p Bonf
6	3	-1.455*	0.378	-3.847	-1.160	0.048
	0	-2.682*	0.600	-4.468	-1.347	0.018
	-3	-4.773*	0.648	-7.366	-2.221	< .001
	-6	-6.136*	0.622	-9.867	-2.975	< .001
	-10	-6.864*	0.568	-12.074	-3.641	< .001
3	0	-1.227	0.401	-3.061	-0.923	0.180
	-3	-3.318*	0.577	-5.750	-1.734	0.003
	-6	-4.682*	0.549	-8.530	-2.572	< .001
	-10	-5.409*	0.504	-10.739	-3.238	< .001
0	-3	-2.091*	0.517	-4.044	-1.219	0.035
	-6	-3.455*	0.627	-5.511	-1.662	0.004
	-10	-4.182*	0.633	-6.602	-1.991	< .001
-3	-6	-1.364*	0.394	-3.464	-1.044	0.091
	-10	-2.091*	0.436	-4.796	-1.446	0.011
-6	-10	-0.727	0.195	-3.730	-1.125	0.059

Note. Cohen's d does not correct for multiple comparisons.

IV. Correlation analysis: Speech perception and listening effort

Pearson correlations were run to describe the relationship between speech perception and listening effort measures. The results are presented in Table. 6. The speech perception was significantly correlated with working memory, peak pupil dilation change and subjective rating of listening and recall effort ($p<0.05$). When speech perception increased the listening effort reduced. Among the working memory measures, working memory was significantly correlated to subjective rating of listening effort. Also, subjective rating of recall effort was significantly correlated to subjective rating of listening effort.

Table 6 Correlation between speech perception and listening effort measures

		Speech recognition	**Working Memory**	**WM Difference**	**Peak Pupil Dilation**	**Listening effort**	**Recall effort**
Speech recognition	r	—					
	p-value	—					
Working memory	r	0.957*	—				
	p-value	< .001	—				
WM Difference	r	0.099	0.111	—			
	p-value	0.257	0.204	—			
Peak Pupil Dilation	r	-0.341*	-0.304*	-0.119	—		
	p-value	< .001	0.002	0.245	—		
Listening effort	r	-0.816*	-0.791*	-0.018	0.132	—	
	p-value	< .001	< .001	0.833	0.195	—	
Recall effort	r	-0.237*	-0.170	0.124	0.043	0.444*	—
	p-value	0.006	0.052	0.155	0.676	< .001	—

*p value is <0.05

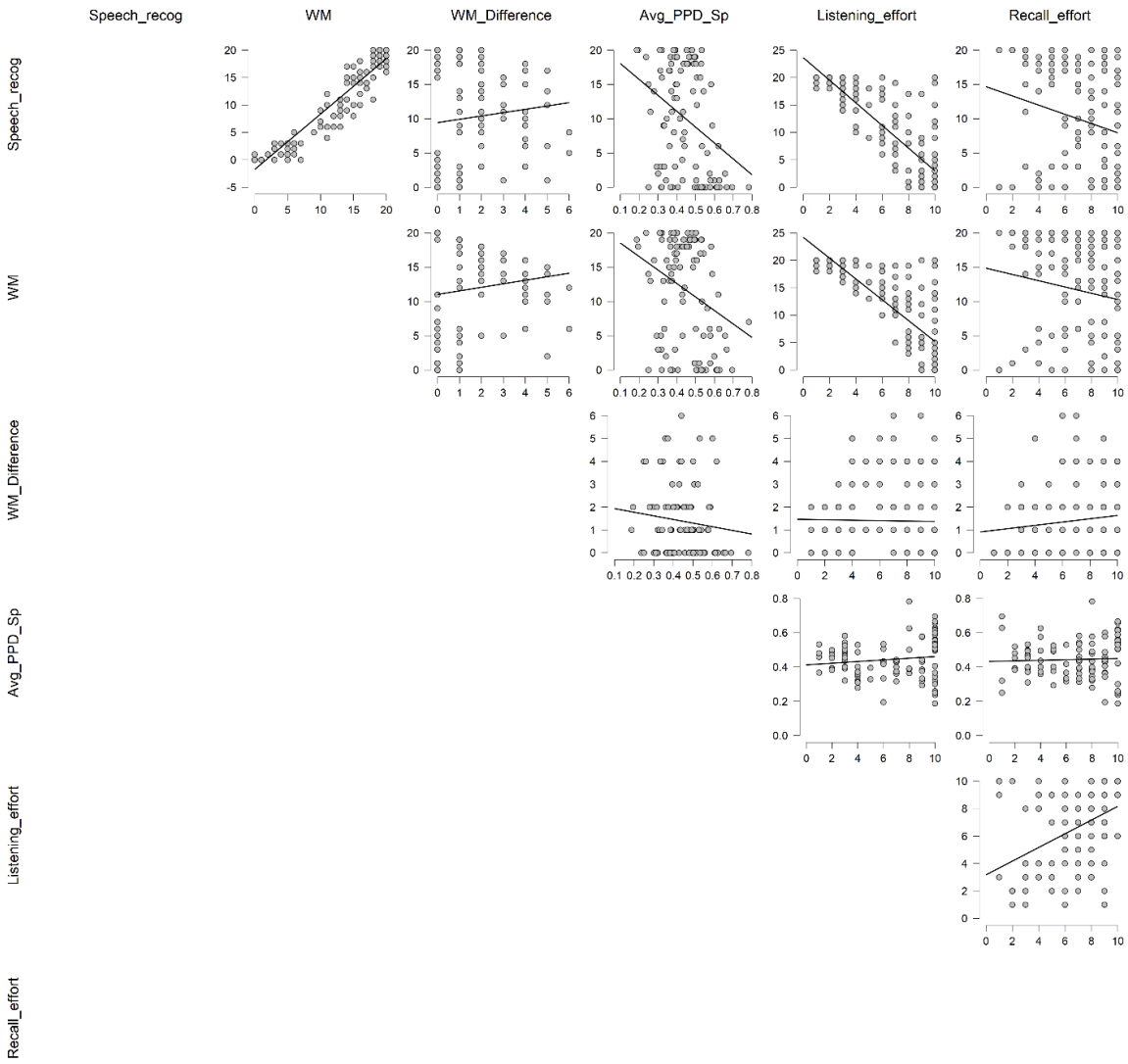

Figure 8 Correlation plot of listening effort measures and speech recognition score. Sp_recog= Speech recognition scores, WM= Working memory, WM_Difference= Working memory difference, Avg_PPD_Sp= Peak pupil dilation change, Listening_effort = Subjective

DISCUSSION

Pupillometry

The mean pupil dilation change observed in the current study ranged between 0.2 mm to 0.6 mm. The magnitude of pupil dilation changes relative to baseline observed in literature ranges below 0.55 mm during sentence recognition task (Wendt, Koelewijn, Książek, Kramer, & Lunner, 2018; Zekveld & Kramer, 2014). The pupillary response is usually measured in a single task paradigm. In the present study the pupil dilation change was measured in a complex task compared to speech recognition. The participants were expected to listen and repeat the sentences. At the same time, they were expected to remember the last word of the sentence. The increased pupil size may be because of the complex task (Padilla, Castro, Quinan, Ruginski, & Creem-regehr, 2020; Piquado, Isaacowitz, & Wingfield, 2010). A study by Padilla et al. (2020) shows the pupil dilation is larger for dual task paradigm compared to single task paradigm. Similarly, Piquado et al. (2010) demonstrated an increase in pupil dilation with increase in memory load. Hence, the difference in the absolute pupil dilation can be attributed to the complex task used in the present study.

Effect of SNR and presentation level

The repeated measures ANOVA did not show any significant effect of SNR or presentation level on pupil dilation when administered on the data set from the five participants with complete data in all conditions. But when missing data were substituted with mean of the condition there was a significant interaction between SNR and

presentation level indicating increase in the power of the study with a greater number of participants. However, the significant result should be interpreted with caution as substitution of data leads to less within group variance and inflation in type I error. The pupil response and speech recognition scores had small significant negative correlation (r= -0.34, p<0.05).

The pupil response increased gradually with decrease in speech perception scores and SNR till -6 dB SNR. The response then dropped at -10 dB SNR indicating disengagement from the task (Zekveld et al., 2014). The trend in pupil response was similar to the trend found by Ohlenforst et al. (2017), Wendt et al. (2018) and Zekveld et al. (2014). Ohlenforst et al. (2017) examined the effect of SNR on pupil dilation with single talker and stationery masker in normal hearing participants with mean age 47 years (SD=12.1). The SNRs used for stationery masker ranged between -12 dB to +16 dB. Wendt et al. (2018) examined the effect of SNR (-20 to +8 dB) on pupil dilation in normal hearing older adults with mean age 65.7 years. The maximum listening effort or maximum pupil dilation in these studies are at an SNR where speech recognition corresponds to 40-80% scores. In the present study the maximum effort is seen when speech recognition is close to 0%. One possible reason for the discrepancy seen in the speech recognition scores at maximum effort or pupil dilation change is age (Peelle, 2018). The Ohlenforst et al. (2017) study shows maximum effort for stationery noise around 40% speech recognition score and in Wendt et al., study the speech recognition at maximum effort condition is 80%. This shows the speech recognition scores at maximum effort point increases with increase in age. In contrast, the speech recognition scores at maximum effort point is around 50% in a study conducted by Zekveld et al. (2014) with

young normal hearing individuals. Zekveld et al. (2014) used single talker masker to create speech in noise conditions. The pupil dilation change function across SNRs is different for single talker masker compared to stationery masker. Single talker maskers show broader range of SNRs with maximum pupil dilation change whereas, stationery masker shows a narrow peak. Hence, the discrepancy in the speech recognition score at maximum effort point may be also due to difference in the stimulus characteristics such as the method used to create stimuli at different signal to noise ratios.

The pupil dilation change in the present study was larger for 65 dB presentation level compared to 50 dB presentation level indicating higher effort at higher presentation level. The difference in pupil dilation was the largest and reached significance only at -6 dB SNR. The effort difference because of presentation level can be attributed to the increased cognitive load and emotional response to increased stimulus redundancy and task difficulty at higher presentation level. There are no studies which examine the effect of overall presentation level on pupil dilation during speech recognition task. A study by Zekveld, Kramer and Festen (2010) examined the effect of background noise level on baseline pupil dilation during a speech recognition task and found no significant effect of level though the magnitude increased with noise level. The intensity of the noise varied between 55 to 63 dB SPL and the results showed no significant effect of noise level on pupil response. In contrast, studies which examined the effect of level on broadband noise perception (no active response), tone or noise detection have shown increase in the pupil response with increase in the level (Antikainen & Niemi, 1983; Bala, Whitchurch, & Takahashi, 2020; Nunnally, Knott, & Duchnowski, 1967). In the studies by Antikainen and Nieme (1983) and Nunnally et al. (1967) the participants were not expected to

actively respond to the stimulus. In the study by Bala, Whitchurch, and Takahashi. (2020), the participants were expected to respond to stimulus by pressing a button. In all the studies the effect of presentation level was examined on the "tonic" pupil size, which is the sustained and absolute pupil dilation in response to stimulus. Tonic pupil dilation is considered to represent the arousal of the person (Peysakhovich, Vachon, & Dehais, 2017).

In the current study the observation of higher effort or larger pupil response to 65 dB (louder presentation level) compared to 50 dB is consistent with the observations of the studies which use non-speech stimulus. However, the response analyzed is the "phasic" pupil dilation which is the transient change in the pupil dilation relative to baseline and represents the cognitive or emotional response to stimulus. The responses were baseline corrected and thus controlled for any arousal or anticipatory effects (Zekveld, Kramer, & Festen, 2010). However, the studies which examined the effect on non-speech stimuli measured average pupil dilation or area under the curve within the first three seconds of the stimulus onset (Antikainen & Niemi, 1983; Bala et al., 2020; Nunnally et al., 1967). Also, Antikinen and Nieme (1983) reported that the pupil dilation decreases as the time increases relative to stimulus onset showing adaptation. Hence, the level effect seen in these studies may represent change in arousal in response to stimulus onset in contrast to cognitive load or emotional response.

To understand the contribution of cognitive load and emotional response to increased pupil dilation, the presentation level was correlated with subjective rating of frustration and disengagement using point-biserial correlation. A positive correlation between frustration and presentation level was hypothesized to represent increased

emotional response at higher presentation level. Similarly, negative correlation between disengagement and presentation level was hypothesized to represent increased task engagement at higher presentation level. The results revealed a significant negative correlation (r = -0.30, p<0.05) between presentation level and disengagement indicating the increased effort at 65 dB is due to increased task engagement. There was no significant relationship between presentation level and frustration level. Hence, it can be argued that the presentation level effect seen with pupil dilation is primarily due to increased task engagement. We postulate that the increased engagement is the result of increased speech redundancy at the higher presentation level. However, this hypothesis needs to be tested by measuring Speech Intelligibility Index or similar measures.

Working memory

Listening effort was measured using two working memory parameters. Working memory represented the number of last words correctly recalled per condition. Working memory difference represented the difference between the possible number of correct recalls and the number of correct recalls per condition. In other words, working memory difference represented the memory cost caused because of noise interference on rehearsal and encoding process of speech. The working memory had a strong significant correlation with speech perception scores (r= 0.96, p<0.001). The working memory difference was not significantly related to speech recognition scores.

Effect of SNR and presentation level

The working memory reduced with reduction in SNR indicating increased effort with reduction in speech recognition scores. The finding is consistent with previous research which showed reduced recall scores or working memory while listening to speech in the presence of noise (Guijo & Horiuti, 2019; Johnson, Xu, Cox, & Pendergrafta, 2015; Pichora-Fuller, Schneider, & Daneman, 1995; Lunner et al., 2016; Ng, 2013; Sarampalis, Kalluri, Edwards, & Hafter, 2009; Strand, Brown, Merchant, Brown, & Smith, 2018). The working memory scores significantly differentiated SNRs from each other except at the ceiling (+6 dB and +3 dB SNR) and floor conditions (-6 dB and -10 dB SNR).

The reduction in working memory is attributed to reduction in encoding of perceived information in memory as more cognitive resources are spent towards understanding degraded speech. Sarampalis, et al. (2009) used +2 dB and -2 dB SNR and found working memory to reduce parallel to speech recognition scores. Similarly, Pichora-Fuller, Schneider, and Daneman (1995) showed working memory reduction corresponding to reduction in speech perception scores. In contrast, Ng (2013) and Lunner et al. (2016) showed reduction in recall task in the absence of speech recognition change. In the present study there was a strong correlation between speech recognition scores and working memory. Hence, another possible reason for working memory reduction when there is concurrent reduction in speech recognition scores is speech intelligibility. When lesser number of sentences are available due to poor SNR this may lead to poor recall scores as there are not many words available to remember.

The effect of poor intelligibility on working memory can confound the effect of increased cognitive load which reduces the encoding of words in the memory. To separate the effect of reduced encoding of words in memory and poor intelligibility on working memory scores, working memory difference was measured. Working memory difference measured the number of recall misses from the number of possible answers indicating the memory cost inflicted by speech perception in noise. The working memory difference increased with reduction in SNR up to 0 dB and -3 dB SNR and then reduced at very poor SNR conditions indicating maximum listening effort or cognitive load when speech recognition scores were in the range of 40-75%. A regression analysis revealed a significant quadratic relationship between speech perception scores and working memory difference (r=0.272, p=0.007). However, speech recognition scores explained only 0.07% variance in working memory difference. Thus, working memory difference can be considered as a measure which shows the cognitive load on encoding words into memory while listening to speech in noise.

The working memory difference was significantly different between 0 and -3 dB SNRs and +6 dB and -10 dB SNR. At other SNRs it was not significantly different. The reduced sensitivity of working memory difference in showing SNR effect compared to working memory score can be due to task difficulty. Pichora-Fuller, Schneider, and Daneman (1995) used different block sizes for recall, varying between two-word recall to eight-word recall and found increased memory cost with increase in the block size. Lunner, et al. (2016), and Ng (2013) used eight sentences in each block and found reduction in working memory score even when speech intelligibility was kept constant. The Lunner, et al. (2016) tested effect of digital noise reduction algorithms and Ng

(2013) compared recall in quite condition to recall in noisy condition (mean +4.1 dB SNR, *SD*= 1.9). Johnson et al. (2015) used five sentences in each block, similar to present study and did not find significant change in working memory scores even when speech recognition changed significantly between SNRs (2, 0, -2 and -4 dB). Hence, increasing the block size may help to improve the sensitivity of working memory difference measure. Another solution is to use low probability sentences as they result in significantly higher memory cost compared to high probability sentences (Pichora-Fuller et al., 1995; Strand et al., 2018).

There was no significant effect of presentation level on both working memory and working memory difference scores. In a study by Amichetti, Stanley, White, and Wingfield (2014), authors used interruption-and-recall (IAR) task in young normal hearing individuals. During the task participants listened to incoming speech information and recalled the words when they perceived they no longer can remember new information. The authors hypothesized reduction in sound level would increase the processing load required to understand and memorize the oncoming information. The words were presented at 25 dB SL and 10 dB SL relative to their SRT and without background noise. The results showed significant reduction in working memory or words recalled with reduction in sound level. In the present study we did not find presentation level effect because of the small block size. In the Amichetti et al. (2014) study, the participants remembered minimum 8 words per trial. However, we cannot rule out the possibility of no memory cost in the SL range used in the current study. Hence there is a need to reevaluate the presentation level effect of working memory cost using more difficult tasks and across a wide range of SLs.

Subjective rating of listening effort

A modified NASA-TLX questionnaire was used to measure self-reported listening effort. The questionnaire included rating of listening effort defined as the effort to listen to and understand the sentences and rating of recall effort defined as the effort to remember and recall the words. The participants were also asked to rate the frustration or irritation experienced (frustration score), how often they gave up listening (disengagement) and their performance level following each experimental condition.

Effect of SNR and presentation level

The subjective rating of listening effort increased monotonically with decrease in SNR. The results agree with previous research which shows increase in subjective rating of listening effort with reduction in SNR and speech recognition scores (Alhanbali et al., 2017; Krueger, Schulte, Brand, & Holube, 2018; Krueger et al., 2017; Strand et al., 2018; Wu, Stangl, Zhang, Perkins, & Eilers, 2016; Zekveld & Kramer, 2014; Zekveld et al., 2010). The function between SNR, speech recognition scores and subjective rating effort varied between studies. Zekveld and Kramer, (2014) measured subjective rating of listening effort at four intelligibility levels ranging between 0 to 100% in young normal hearing individuals. Krueger et al. (2017) and Krueger et al. (2018) measured subjective rating of listening effort using adaptive procedure across a wide range of SNRs (-24 to +12 dB SNR) in both normal hearing- and hearing-impaired individuals. The results from these studies showed a linear trend of subjective rating of listening effort, whereas, Wu et al. (2016) showed a non-linear trend in a study measuring subjective listening effort during a dual-task paradigm, where SNRs varied between +10 to -10dB with reference to

SNR50. The different trends or functions observed may be due to difference in the task used in the study. According to theory of dissociation by Yeh and Wicken (1984), the subjective workload is sensitive to the aggregate of resource investment (Yeh & Wicken, 1984). Hence, it can be argued that the subjective rating during a dual task paradigm is affected by the amount of cognitive resources spent for both understanding speech and performing the secondary task. In the Wu et al. (2016) study, though the listening effort increases in reduction in SNR at very difficult conditions the overall resource allocation reduces due to decrease in intelligibility or need to process information. The reduction in cognitive load at very poor SNR or difficult condition is supported by reduction in reaction time to perform the secondary task. From this observation it can be hypothesized that the reduction in subjective rating at the poor SNRs is due to reduction in overall cognitive load. In contrast, in the present study, the participants were asked rate effort separately for listening and recall tasks. Hence, the trend difference could be because of task difference between the studies. There are no other studies in the literature that explore the relationship between subjective rating scale and SNR during a dual-task paradigm across a wide range of SNRs. Hence, there is a need for more studies which explore this relationship in order to examine this hypothesis.

The subjective rating of recall effort was a new scale introduced in the current study to separate the effect of speech perception in noise and recall task on subjective rating of effort. There was no significant effect of SNR on recall effort. The average data showed a non-linear trend where recall effort increased with reduction in SNR from +6 dB to 0 dB SNR and remained constant across 0 dB SNR to -6 dB SNR and reduced at -10 dB SNR. The trend was similar to trend in working memory difference. However,

there was high inter-subject variability in data resulting reduced power. Both, subjective rating of listening effort and recall effort did not show any significant effect of presentation level. Despite high individual variability, recall effort showed a non-linear relationship with SNR at 65 dB SPL and the effort remained constant across conditions at 50 dB SPL. The reasons for the effect of presentation level on average recall effort ratings is not clear. Use of cognitive interview techniques may facilitate the understanding of strategies used by participants to rate recall effort and warrants further exploration.

The subjective rating of listening effort is influenced by the perceived performance and is the reason for disassociation between objective, behavioral and subjective measures of listening effort (Moore & Picou, 2018). In the present study, there was a strong positive correlation between subjective rating of listening effort and perceived performance (r=0.77, p<0.001). The recall effort also showed a significant moderate positive correlation with perceived performance (r= 0.35, p<0.05). There may be a possible influence of working memory on recall effort in addition to performance (r= -0.17, p = 0.052), resulting in increased variability at extreme SNR conditions (+6- and -10-dB SNR). However, the results warrant more studies due to poor power and small sample size.

Comparative sensitivity of listening effort measures

A Pearson's correlation analysis was run to understand the relationship between listening effort measures and speech recognition scores (Figure 8). The data for each condition from each participant was considered one data point (N=132). As pupil dilation

had missing data for certain conditions, missing data was excluded listwise. The listening effort measures- working memory, peak pupil dilation change and subjective rating of listening and recall effort showed increased listening effort with reduction with speech recognition scores reduction. The influence of intelligibility on listening effort was less for recall effort (r= -0.27, p<0.05) and pupil measures (r= 0.-34, p<0.05) compared to working memory (r= 0.96, p<0.05) and subjective rating of listening effort (r= 0.80, p<0.05). The working memory difference was not related to speech recognition scores.

The peak pupil dilation and working memory had small significant positive correlation indicating higher peak pupil dilation with higher working memory. As both peak pupil dilation and working memory are related to speech recognition scores, a regression analysis was conducted to predict peak pupil dilation by working memory controlling for speech recognition scores. Working memory (p>0.05) was not a significant predictor of peak pupil dilation when controlled for speech recognition scores, indicating peak pupil dilation change and working memory to have different underlying construct while measuring listening effort. Subjective rating of listening effort and working memory had a strong positive correlation (r= 0.79, p<0.05). Similarly, when controlled for speech intelligibility, there was no significant relationship between working memory and subjective rating of listening effort. The subjective rating of listening effort and recall effort were significantly related to each other even after accounting for perceived performance scores and speech recognition scores, indicating common underlying construct for subjective rating measures.

The sensitivity of listening effort measures was compared based on significant effect of SNR and presentation level on the measures and effect sizes from two-way

repeated measures analysis. The results here should be interpreted with caution, as the sample size and power are different across listening effort measures. The peak pupil dilation (with mean substitution), working memory, working memory difference and subjective rating of listening effort showed significant main effect of SNR. Of all the measures, working memory (η^2_p = 0.98) was most sensitive to SNR effect, followed by subjective rating of listening effort (η^2_p = 0.84), working memory difference (η^2_p = 0.52) and peak pupil dilation (η^2_p = 0.40). A study by Seeman and Sim (2015) compared physiological (heart rate, skin conductance), behavioral (reaction time), and subjective measures of listening effort (NASA-TLX) across SNRs ranging between 0 dB to +15 dB SNR in young normal hearing individuals. The results showed subjective measures to be more sensitive compared to physiological and behavioral measure. The behavioral measure was estimated at +5- and +15-dB SNR, both positive SNRs which result in near normal speech recognition scores. Similarly, Johnson et al., (2015), examined the comparative sensitivity of subjective rating scale and working memory (listening span) in young normal hearing individuals and found subjective rating to be more sensitive to SNR changes (+2 to -4 dB) compared to working memory. In the current study the effect of SNR was examined over a large range of SNRs (+6 to -10 dB). As behavioral measure was highly sensitive to intelligibility behavioral measures along with subjective rating of listening effort showed high sensitivity to the effect of SNR unlike studies by Seeman and Sim (2015) and Johnson et al. (2015).

Alhanbali, et al. (2019) compared the sensitivity of subjective (self-reported effort), physiological measures (skin conductance, pupillometry and electroencephalography) in normal and hearing-impaired individuals at SNR

corresponding to 71% intelligibility level. The results showed pupil dilation to be more sensitive compared to other physiological and subjective measure. The possible reason for the discrepancy between Alhanbali et al. (2019) study and current study is the population tested and the SNR conditions. As all listening effort measures other than working memory difference was significantly related to speech intelligibility, there is a need to examine the sensitivity of the paradigm used in our study when controlling for speech intelligibility and also in hearing impaired individuals.

Of all the listening effort measures only, peak pupil dilation showed significant presentation level effect at -6 dB SNR. When effect size was compared peak pupil dilation (η^2_p = 0.27) was more sensitive to presentation level effect followed by working memory (η^2_p = 0.15). Working memory difference and subjective rating of listening effort explained very less variability due to presentation level.

Individual data analysis

The complete five data sets are plotted in Figure 10 for trend analysis. The range of listening effort measures was rescaled to 0-10 units to facilitate comparison. The following formula was used where Y_{adj} was the rescaled value, Y was the observed value, Y_{min} was the minimum value observed in the data, Y_{range} was the range of values observed for every variable.

$$Y_{adj} = \left(\frac{Y - Y_{min}}{Y_{range}}\right) 10$$

In Figure 10, except working memory, for all other variables higher value represented higher listening effort. All participants except Sub 3 and 4 mimic a non-linear

peak pupil dilation function vs SNR like averaged data. Except Sub 3 and 5 all other participants showed relatively higher dilation at 65 dB SPL compared to 50 dB SPL around -6 dB SNR. Working memory data showed consistent negative slope relative to SNR. The magnitude of slope of subjective rating of listening effort changed from subject to subject, however, the relationship between effort and SNR remained consistently positive. The recall effort data either followed the trend of working memory (Sub 1 and 3) or the subjective rating of listening effort (Sub 2, 4 and 5) and more consistently reduced in magnitude at -10 dB SNR compared to listening effort rating. This trend resulted in a non-linear trend of recall effort average data. Of all participants Sub 3 showed less effect of SNR and all listening effort measures show least change with SNR changes. Also, the maximum effort as shown by peak pupil dilation and behavioral method (memory cost) are different and further supporting the notion that both methods have different underlying construct.

The most interesting observation is the interaction point occurring between listening effort measure around -3 dB or -6 dB SNR. The relationship between working memory difference, recall effort with pupil dilation changes from positive to negative. In other words, before -3- or -6-dB SNR the listening effort increased or remained constant as measured by pupil dilation, working memory difference or memory cost and recall effort. Around the intersection point, though pupil dilation showed increase in effort working memory difference and recall effort showed reduction in effort indicating earlier breakdown point for behavioral and subjective measures of listening effort. To better understand the relationship between these measures, a correlation analysis of data at each

SNR condition will help. However, the correlational analysis was not done for the current data owing to small sample size and non-normal data distribution (Appendix 5).

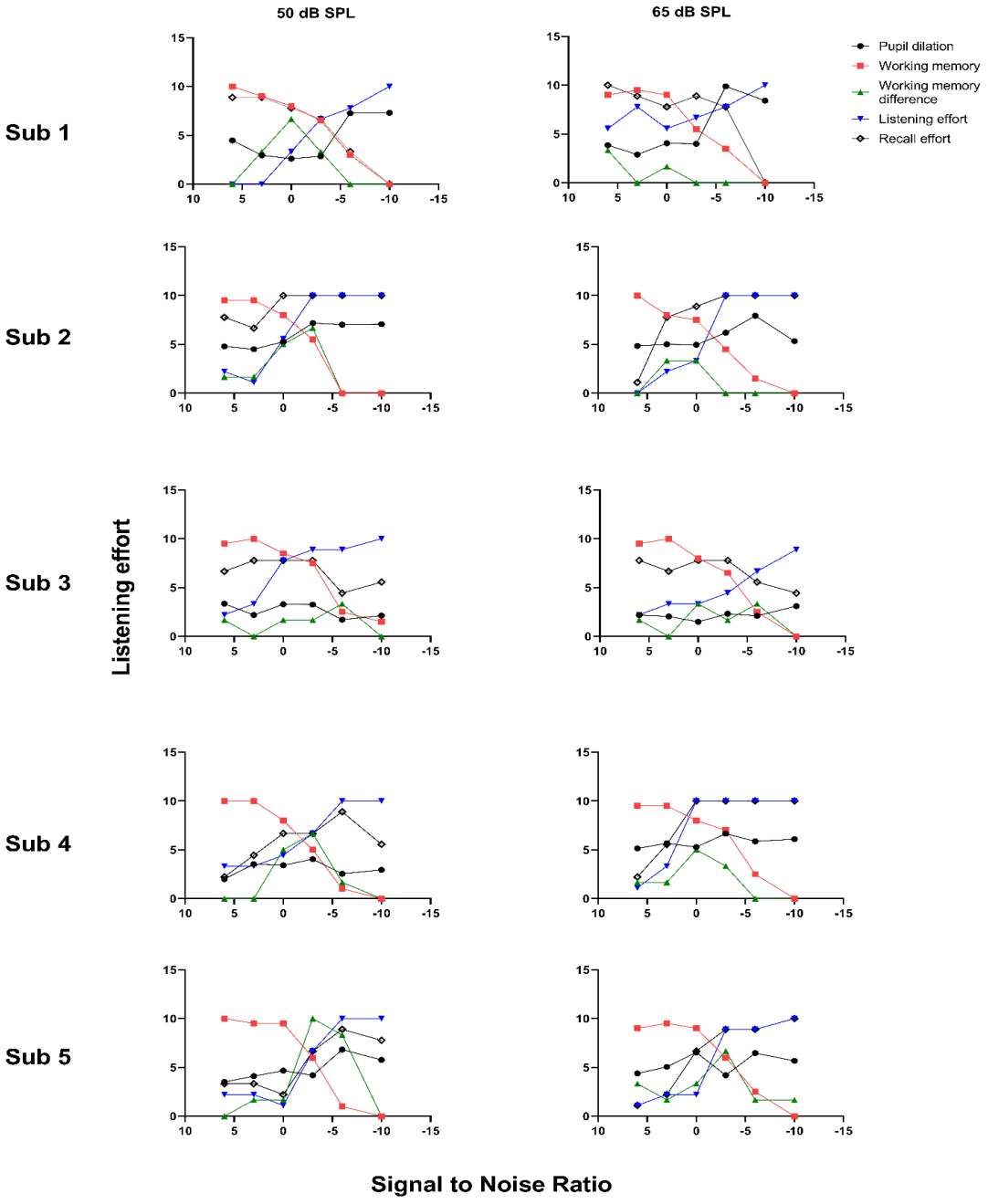

Figure 9 Comparison of listening effort measures in individuals across SNRs and presentation level

Conclusions

The estimation of listening effort with peak pupil dilation, working memory, and subjective rating of listening and recall effort were significantly related to speech intelligibility or recognition scores. When controlled for speech intelligibility all listening effort measures were not significantly related indicating different underlying constructs. All listening effort measures except recall effort showed significant effect of SNR. Working memory was most sensitive to SNR effect, followed by subjective rating of listening effort, working memory difference and peak pupil dilation. Only peak pupil dilation showed significantly higher effort for higher presentation level. As speech intelligibility was a significant factor deciding the listening effort with change in SNR, there is a need to examine the sensitivity of the paradigm used in the present study controlling for speech intelligibility.

Limitations

The study had a smaller sample size compared to sample size estimated with pre-study power analysis. The study sample size could not be met due to COVID-19 restrictions on data collection. Also, there was data loss observed for pupil data due to technical reasons. Out of 11 participants who completed the study, two participants did not have any useful pupil data and three participants had data loss in one out of twelve conditions. Another participant had data loss in seven conditions. The review of video recording of the testing showed loss of data even when participants maintained gaze fixation. The loss of data resulted in unequal sample size for different listening effort measures and hence different power for average data analysis using two-way ANOVA.

APPENDIX-1

REVIEW OF LITERATURE

Hearing and cognition

Hearing ability in humans fulfills the purpose of communication. Hearing loss hinders oral-aural communication by reducing audibility of sounds and also reducing the clarity of sounds (Moore, 1996). There are primarily two views of hearing loss which forms the basis for diagnosis and rehabilitation models. One is site-of-lesion view and the other is processing view (Pichora-Fuller & Singh, 2006). According to site of lesion view, the hearing pathway is considered as a series of units which are overlapping and is considered as a system dominated by afferent nerves. The speech perception is considered basically through bottom-up process, though it considers the influence of efferent nervous system on peripheral hearing. The current diagnosis process and rehabilitation models are primarily influenced by this view, where the perception of simple sounds in ideal conditions are considered as yard sticks of improvements (Pichora-Fuller et al., 2016). In contrast the processing view considers hearing as a combination of bottom-up and top-down processing. This view also considers cognition as an essential part of hearing. Kiessling et al. (2003) described four functions of auditory system. They are as follows,

i. Hearing: The passive perception of auditory stimulus in the surrounding
ii. Listening: The perception of auditory stimulus with attention to stimulus
iii. Comprehension: Unidirectional understanding of the auditory information

iv. Communication: Two-directional exchange of information in auditory mode

When the functioning of auditory system and language processing are considered, communication is an active dynamic process which is just not based on the involvement of the peripheral auditory systems but more central processing. This understanding that aural communication is more complex with the involvement of cognition processing has gained more attention since past two decades and there is argument that involving the cognitive assessment in the process of rehabilitation will be closer to real life experiences. The cognitive processes like memory, processing speed and language are now considered essential part of successful aural-oral communication.

Kahneman's Capacity model postulates a general cognitive framework which helps in the processing of sensory information (Kahneman, 1973). According to the capacity model the cognitive resources are limited in persons and the resource allocation to sensory information decides the behavioral response to stimuli. In line with the theory, several studies have shown reliance of auditory processing on cognitive processes and disruption in cognitive processing due to hearing loss even when audibility is taken care of. The studies that tried to find the factors which predict the variance in speech perception across different subject groups showed a small part of variation to depend on cognitive factor memory. A large-scale study by Humes (2003) on 134 subjects showed verbal intelligence quotient is positively related to speech recognition score and subjective perception of benefit and negatively related to hearing aid use (or uptake). Studies have also shown speech perception in degraded stimulus conditions to interfere with memory (Cousins, Dar, Wingfield, & Miller, 2014; Rabbitt, 1968). Rabbitt (1968) in

his classic study showed reduced memory for words while listening to speech in noise. A Study by Pichora-Fuller and colleagues (1995) showed significant reduction in listening span (a measure of working memory) while listening to speech in noise in both young and older adults with normal hearing abilities. Older adults had significantly lesser working memory (listening span) when compared to younger adults. This shows the processing and storage of information become taxing while listening to speech in noise and especially in older adults. Due to interdependency of cognition and communication, it is proposed that including cognitive assessment in clinical test battery helps to account for individual differences in communication abilities.

Some of the clinical observations which support inclusion of cognitive test in everyday clinical practice are: (1) the high variance in speech perception scores seen in persons with hearing loss despite having similar audiological characteristics such as pure-tone thresholds or when audibility is restored, (Verschuure & Benthem, 1992); (2) complaints from patients about increased listening effort and fatigue regardless of having achieved good audibility and problems in understanding speech at supra-threshold level (Pichora-Fuller, 2010). These observations are supported by research findings which show decreased processing speed, increased processing load as indicated by EEG, fMRI and fNIR measures, reduced listening span, and performance on a secondary task in persons with hearing loss even when speech intelligibility is accounted for (Alhanbali et al., 2019; Wijayasiri, Hartley & Wiggins, 2017; Wild et al., 2012); Differences in cognitive abilities such as working memory are proposed as the reason for individual differences in communication abilities in persons with hearing loss and older individuals when controlled for audibility and speech intelligibility (Rönnberg et al., 2013). Hence,

there is a strong need to include a cognitive measure in the clinical settings to better understand the communication abilities of a person to facilitate choosing suitable rehabilitation options.

Listening effort: a cognitive measure for clinics

Listening effort and fatigue are two concepts which are based on the cognitive models. Listening fatigue is a common complaint of persons with hearing loss. Several studies have shown persons with hearing loss to experience more listening fatigue compared to normal hearing individuals (Alhanbali et al., 2017; Bess & Hornsby, 2014; Hornsby, 2013; Alhanbali et al., 2017; Pichora-Fuller et al., 2015). A study by Nachtegaal et al. (2009) revealed that persons with hearing loss require more recovery time after working compared to persons without hearing loss. Similarly, another study by Kramer et al. (2006) revealed burnout and fatigue due to hearing loss as a reason for increased frequency of sick leaves in persons with hearing loss. Pichora-Fuller et al. (2015) investigated the effect of hearing loss on listening effort, quality of social interaction and social isolation and found higher listening effort, reduced quality of social interaction and increased social isolation in persons with hearing loss. The increased listening fatigue in persons with hearing loss is due to use of increased listening effort for an extended period of times (Pichora-Fuller et al., 2016). As listening fatigue is a complex variable, listening effort is used as an alternative measure.

Listening effort is defined as the mental effort experienced due to deliberate allocation of mental/cognitive resources to overcome obstacles in the goal pursuit while involved in a listening task (Pichora-Fuller et al., 2016). Listening effort can be proposed

as a suitable clinical cognitive tool as it represents the cognitive resource used in the process of speech understanding and thus better explain the individual differences even when intelligibility and audibility factors are accounted.

Until 2015 there was no consensus on the definition of listening effort, or the terminology to represent the same. In the Eriksholm workshop in 2015 researchers from different disciplines came together to address the issues such as lack of consistent definition in literature and lack of theoretical model. The evidence collected so far was evaluated to come to consensus with the definition of listening effort. During the workshop listening effort was defined as "mental effort experienced due to deliberate allocation of mental resources to overcome obstacles in the goal pursuit while involved in a listening task" (Pichora-Fuller et al., 2016). At the workshop a framework for understanding the mechanism of listening effort was also formulated. This framework majorly borrowed the concepts from Kahneman's Capacity Attention model (Kahneman, 1973). Further evidence based on other cognitive theories such as attention, processing speed, socio-cognitive models, physiological motivation and arousal theories, ease of language understanding theory were discussed, and further components based on these theories were incorporated in the framework.

According to this framework, the various task demands results in the arousal of the sympathetic nervous system resulting in physiological responses such as pupil dilation, increase in skin conductance and increased cardiac response. Five factors were considered important in creating the task demand. They are source factors (example: new/unknown accent), transmission factors (example: noise, reverberation in the room), listener factors (example: hearing loss, reduced cognitive capacity), message factors

(example: vocabulary, semantic knowledge), and context factors (example: knowledge of the communication set-up). These task factors are assumed to increase the listening effort. Once the demand results in the arousal of the sympathetic system the person engages in cost-benefit analysis (evaluation of demand on cost) based on the activities he need to get involved. This analysis takes place before allocating the mental resources to engage in the task. Based on the cost-benefit analysis if the person feels there is benefit in engaging in the activity then s/he allocates mental resources in the activity. However, this evaluation process can also be influenced by other factors such as fatigue, low arousal, and (dis)pleasure. If a person is experiencing fatigue, low arousal or if s/he is not deriving pleasure by involving in the activity then that person may decide to quit participating in the activity. Similarly, the allocation policy which decides to what extent mental resources should be used for the activity can get affected by the kind of attention that activity involves. For example, if it is automatic attention (example: response to name call) the allocation policy may expend less mental resources for the activity. In comparison, if the person is purposefully attending to an activity (example: to a particular person's voice) then s/he may expend more mental resources for the activity. Hence, factors like fatigue, low arousal and (dis)pleasure may in turn affect intended attention and result in changes in allocation policy.

Once the person starts engaging in the activity, following the directions of the allocation policy, this may result in physiological or behavioral responses. Four types of responses are explained under this model. They are cognitive-behavioral responses (example: recall, dual task paradigm response cost), arousal responses (example: pupil dilation, skin conductance, cardiac response), brain (electrophysiological responses,

neural imaging), and self-report responses. These responses are proposed as indicators to measure listening effort.

Importance of measuring Listening effort

Speech perception measures and listening effort

The listening effort measure has been reported to be a more sensitive measure compared to speech perception tests (Sarampalis et al., 2009; Winn, Edwards, & Litovsky, 2016). several studies have shown listening effort measure to be more sensitive while investigating aspects like the effect of aging, benefit of hearing aid algorithms, benefits associated with cochlear implants compared to speech intelligibility. Gosseline and Gagne (2011a, 2011b) investigated the effect of age on listening effort using a dual task paradigm when speech recognition scores were equalized. In both studies 25 subjects with normal hearing participated in each group (young vs older). The studies involved a tactile pattern recognition task as secondary task and the response cost was measured between single task and dual task. Both studies revealed older individuals to have higher response cost in terms of pattern recognition accuracy and response time compared to younger individuals even when both groups had equivalent speech recognition scores. This shows listening effort as a sensitive measure in understanding the effect of age compared to speech intelligibility measure.

A study by Sarampalis et al. (2009) showed listening effort to be sensitive in measuring the benefit of digital noise reduction (DNR) compared to speech intelligibility. The authors found no difference between DNR-on, off condition for speech intelligibility measures; however, there was significant difference between the two conditions in terms

of listening effort. Another study by Johnson et al. (2016) evaluated the difference between premier level hearing aid and basic level hearing aid in terms of speech intelligibility and listening effort. The results showed for one manufacturer listening effort measure did depict the benefit of premier hearing aid. In addition, a study has shown persons with better working memory to get benefitted from fast compression compared to those with poor working memory. Thus, it can be hypothesized that listening effort is a more sensitive measure to evaluate the candidacy for different algorithms and devices.

Similarly, studies by Pals et al. (2013), Winn (2016) and Winn, Edwards, and Litovsky (2016) showed listening effort as a sensitive measure to detect the effect of spectral distortion compared to speech intelligibility measure. In their studies the authors provided spectrally degraded (vocoded speech) stimulus to individuals and investigated the rate of change in speech intelligibility and listening effort as measured with dual task paradigm (Pals et al., 2013) and pupillometry (Winn et al., 2016) across different number of channels. The authors found speech intelligibility to plateau after six to eight channels; however, the listening effort did improve even at higher number of electrodes. Thus, it can be assumed that listening effort is a more sensitive measure compared to speech intelligibility in certain aspects during cochlear implant programming.

Listening effort explores multiple dimensions of auditory stimulus perception

According the FUEL framework, listening effort is deliberate allocation of mental/cognitive resources to complete a listening task (Pichora-Fuller, 2016). Assessing at what cost a person achieved a certain performance level can indicate to what degree a

person relies on cognitive resources during a listening task (Edwards, 2007). While assessing listening effort, along with understanding the effect of listening condition on cognitive resource allocation and its consequences on speech perception, we can also get information on how attention, general mental status of the person, motivation affects speech perception. Motivation is considered as an important modulator of effort whose mobilization can affect long term fatigue (Richter, Gendolla, & Wright, 2016). A study conducted by Richter (2016) measured listening effort using cardio-vascular reactivity as an index during an auditory discrimination task. The results showed greater listening effort when there was greater success importance, manipulated using monetary rewards in high listening demand condition compared to when listening demand was low. The study conducted by Koelewijn, Zekveld, Lunner, and Kramer (2018) showed higher listening effort as measured with pupillometry for high reward condition to low reward condition. These results are consistent with the FUEL framework, which states cost-benefit analysis to affect listening effort.

Furthermore, studies have shown listening effort to indicate the level of engagement is a given task. The pupillometry studies done to explore the effect of SNR show decrease in the pupil dilation in very difficult speech perception conditions (sentence recognition scores less than 30%) (Koelewijn, Kluiver, Shinn-cunningham, Adriana, & Kramer, 2015; Wendt, Koelewijn, Książek, Kramer, & Lunner, 2018; Zekveld, Kramer, & Festen, 2010). The authors attribute this decrease in pupil dilation (or reduction in listening effort) to disengagement from task. Listening effort also affected by attention. Various studies have shown increased pupil diameter during intentional active listening compared to passive listening (Laeng, Eidet, Sulutvedt, &

Panksepp, 2016). According to the FUEL framework, automatic attentional and intentional attention are effective modulators of listening effort. Thus, as listening effort explores different dimensions of auditory stimulus perception, measuring listening measure may be useful in examining the interindividual differences in speech perception. Also, sensitivity of listening effort to multiple internal factors makes listening effort more ecologically valid measure.

Listening effort assesses different levels and processes of auditory system

Listening effort measures help to assess the top-down processing of speech. The top-down processing or use of cognitive resources is useful while listening in adverse listening conditions. Even normal hearing individuals recruit working memory resources when there is degradation in the phonological information of speech (Rönnberg, Holmer, & Rudner, 2019). When we measure the effect of task-load on listening effort for example, speech perception in the presence of background noise or perception of speech by non-native speakers of language and internal factors like presence of hearing loss, listening effort reflects the cost of resource consumption by bottom-up process on top-down processing. This feature of listening effort can help to explain the interindividual differences in speech perception.

Though there is uncertainty, there is accumulating evidence to show listening effort as a sensitive measure in deciding candidacy for persons with hearing loss, to evaluate the effect of different populations and to evaluate the outcome of intervention strategies compared to speech intelligibility measure. In addition, as listening effort is a tool which explains the suprathreshold speech perception variance this may provide a

holistic view on the problems of the person with hearing loss and may help to individualize and improve the effectiveness of rehabilitation programs.

Methods of measuring Listening Effort

Listening effort is the mental effort exerted to get involved in the listening task. There are different kinds of measures used to measure listening effort. They can be classified into three categories. They are as follows,

a. Cognitive-behavioral methods: These methods are based on the assumption that listening effort expended by the person during a listening task results in changes in behavior of interest. The behavior of interest can be a listener's performance on a secondary task in a dual-task paradigm, updating or inhibition behaviors when involved in working memory tasks etc..
b. Behavioral methods: These methods are based on the assumption that listening effort expended by the person during a listening task results in changes in behavior of interest. The behavior of interest can be secondary task performance in dual-task paradigm, updating or inhibition behaviors when involved in working memory tasks etc. These methods are based on the theory of limited capacity (Kahneman, 1975). According to this theory, when a person performs two activities simultaneously the cognitive resources are said to be distributed between two activities based on the importance of the task as there is only limited amount of resources available. In dual-task paradigm when a person is asked to prioritize speech recognition task (or primary task), this will result in performance decrement in secondary task (behavior of interest). This reduction in performance or response cost is considered as an indicator of listening effort.

In case working memory, the same principle applies; however, in this task when a person involves in a difficult listening situation the limited cognitive resources are utilized to understand the speech stimulus and this affects the ability to store that information resulting in poor memory (behavior of interest). Behavioral method is considered as an objective test as there is a provision for reliable quantification of the responses.

c. Physiological methods: This follows the principle that when a person exerts mental effort, it results in physiological changes due to the activation of central nervous system circuits such as sympathetic nervous system (for example, pupil dilation, increase in skin conductance and increase in heart rate). The other hypothesis which plays a role in physiological measures is the change in activation patterns of the brain when person is experiencing an increase in listening effort (for example, changes in activation in the central nervous system-frontal cortex, cingulate opercula region etc.) (Strand, Brown, Merchant, Brown, & Smith, 2018). Electrophysiological tests (MMN, P300, N2b etc.), magnetoencephalography, neuro-imaging methods like fMRI are used to assess the brain activity. Pupillometry is found to be more sensitive tool in the measurement of listening effort compared to increased skin conductance and heart rate with increase in mental effort or stress (Strand et al., 2018). Changes in salivary cortisol level is also considered as a physiological indicator stress due to changes in listening effort.

d. Subjective methods: Subjective self-report methods rely on the direct expression of subject's experience. This can include procedures where participants rate the

amount of effort perceived following a speech perception task or it can involve rating scale which measures effort experienced generalized to a day. There are no standardized subjective scales are available currently. The commonly used scale is the sub-section of Speech, Spatial and Qualities questionnaire (SSQ) (Gatehouse, & Noble, 2004). The other common measures used are the one-dimensional questions (single questions) which require patients to rate the amount of effort experienced following a speech perception task.

Cognitive-behavioral methods

According to FUEL, cognitive-behavioral methods can be used to understand the effect of task demands on listening effort. That is the effect stimulus related factors (SNR, accent, lexical context etc.), subject related factors (like, age, hearing loss, cognitive ability etc.) have on listening effort. Several behavioral tests have been reported in the literature to measure listening effort. They can be broadly classified into two categories (figure 10).

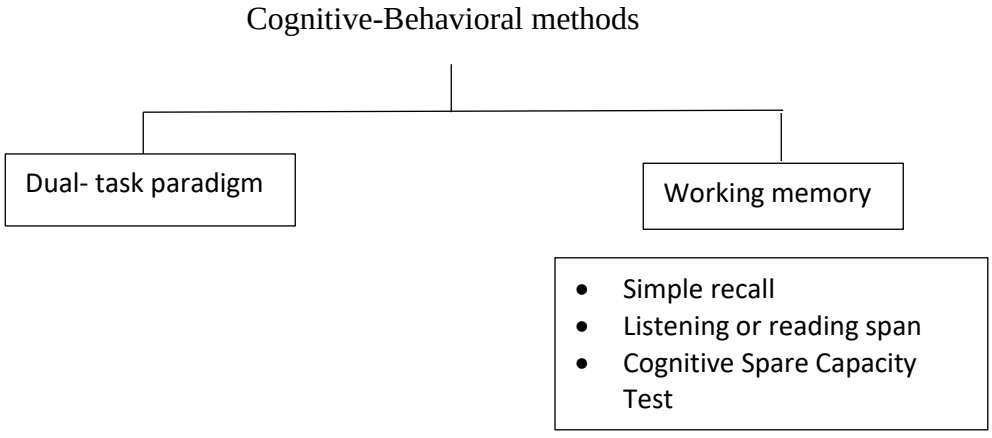

Figure 10 Flow chart of types of cognitive-behavioral methods of listening effort measurement

Dual task paradigms

Dual task paradigms are the most common and widely used cognitive-behavioral method in listening effort measurement. These methods are based on the cognitive resource theory or limited capacity theory proposed by Kahneman (1973). According to this theory every person will have limited cognitive resource and it is allocated to different tasks based on the importance of the task. If a person is required to participate in more than one task the resource gets divided between the tasks and if maintaining the performance in one of the tasks is important (primary task), then that task gets the major share of the resource or it dominates compared to the less important task. This difference in resource allocation can reduce the performance of the less important task or the secondary task when compared to its performance in the absence of primary task. Based on this concept, the dual task paradigm was designed where initially the person's performance on the primary task and a secondary task will be measured individually (or in single task condition). Later the person would be asked to participate in primary and secondary tasks simultaneously or sequentially and his or her performance will be measured in dual task condition. The secondary task's performance difference between single and dual task condition is named response cost or dual task response cost. The magnitude of this response cost is considered as an indication of listening effort. This method conforms with ease of listening hypothesis which states better cognitive capacity reduces processing load in difficult conditions (Van Der Meer et al., 2010).

Methodological variations in dual-task paradigm

Different types of dual task paradigms are reported in the literature to measure listening effort. The following are the methodological differences that are found between studies and their effect on the results.

a. *Dual task:* The dual-task test can be administered in two different types. One is concurrent presentation, the other is sequential. In concurrent method, the subject will be asked to involve in the primary task of speech recognition and simultaneously s/he will be asked to perform the secondary task. For example, in the study conducted by Desjardins and Doherty (2014), the subjects were required to engage in the primary task of speech recognition and at the same time they were asked to follow the digits that appeared on the screen with the help of the mouse. In concurrent task it is assumed that the method is more ecologically valid as in real life persons are required to engage in multi-tasking. Also, concurrent task is assumed to be more cognitively tasking compared to simple recall involved in the sequential task.

   In contrast, a sequential task will require the subject to perform primary task and following the primary task perform the secondary task. However, the stimulus processing of primary and secondary task occurs simultaneously. The study conducted by Rakerd, Seitz and Whearty (1996) employed a sequential task. In this study the participants were asked to perform primary task of speech recognition and during this task they were presented with strings of number. Following the response for speech recognition task the participants were asked to recall the numbers presented before (Rakerd, Seitz, & Whearty, 1996). Though both the methods can be used, in literature there seems to be a strong bias for concurrent task as majority

of studies use concurrent procedure. As mentioned above, reason for this could be the assumptions about cognitive load and ecological validity (Gagné, Besser, & Lemke, 2017).

b. *Primary task related factors:* Primary task related factors which differ across studies are as follows,

i. Material used: There is a wide variation in the test materials used in the primary task. Majority of the studies use sentence recognition test. Other than sentence recognition tests there are instances where studies use syllable recognition, word recognition (Picou & Ricketts, 2014a), passage recognition tests. Though there is wide variation in the use of speech materials, there is no clear evidence to show preferable material for primary task (Gagné et al., 2017).

ii. Signal-to-noise ratio (SNR): The signal-to-noise ratio use in the study is shown to affect the sensitivity of the dual-task paradigm. Studies have shown listening effort to decrease with increase in SNR. However, it is important to notice that this decrease is also accompanied with increase in the primary task performance. This indicates that listening effort reduces with increase in audibility or speech recognition performance. A study by Wu et al. (2014) examined the effect of different SNRs on the dual task response cost. The results revealed a non-linear pattern in reaction-time responses with changing SNR. The reaction time was longest for the SNRs which resulted in primary speech recognition scores within 30-50% range. The reaction time reduced for SNRs which resulted in response lesser than 30% or greater than 50% response range. Here the reduction in reaction time with reduction in speech recognition performance below 30% of response

range is in contrary with the results of the previous studies. However, this can be attributed to the phenomenon of quitting the process of hearing when condition is very difficult (Picou & Ricketts, 2014a). Thus, it is important to select the level of performance at which listening effort test to be conducted. In a recent study by Strand et al. (2018), where authors examined the convergent validity of different listening tests, 50% and 80% performance levels were considered to avoid the effect of ceiling and floor. That is the authors consider performance levels that most probably brings a change in reaction time when compared to baseline.

iii. Linguistic context: Studies have used speech material with different linguistic load in the primary task and results of these studies reveal low-predictable material to result in lesser listening effort compared to high-predictable sentence (Pichora-Fuller, Schneider, & Daneman, 1995).

c. *Secondary task related factors:* The secondary task related factors such as the type of task, the outcome measures used, the metric used for measurement can have effect on the results.

i. Type of secondary task: There is a wide variation in the type of secondary task used in the studies. There seems to be a common assumption, that there is no effect of type of secondary task on the results (Gagné et al., 2017). Visual pattern recognition, tactile pattern recognition (Gosselin & Gagne, 2011a, 2011b), simple visual probe (Picou & Ricketts, 2014b), complex visual probe (Picou & Ricketts, 2014b; Strand et al., 2018), semantic judgements (Picou & Ricketts, 2014b; Strand et al., 2018), syntactic judgement, car driving simulation (Wu et al., 2014), visual motor tracking (Desjardins, 2016; Desjardins & Doherty, 2014) etc., are some

examples of different secondary tasks employed in the dual-task paradigm. Picou and Ricketts, (2014b) examined the effect of type of secondary task utilized on listening effort outcomes. They conducted two experiments in which the participants were asked to engage in simple visual probe, complex visual probe and category recognition of the noun (the words presented for primary task) secondary tasks. In first experiment normal hearing individuals participated in the study and in the second experiment persons with hearing impairment participated in the study. The results revealed category recognition of the noun to be the only sensitive secondary task to measure listening effort for both the subject groups. The authors propose that the reason could be because of the deeper processing required in category recognition task as it involves linguistic processing (semantic judgement: recognizing whether word is noun or verb) required for the primary task. In contrast the study conducted by Strand et al. (2018) with similar procedure as that of Picou and Ricketts (2014b) in normal hearing individuals show that both complex visual probe and category recognition of nouns are sensitive to measure listening effort. However, the results of the study were in agreement with the Picou and Ricketts, (2014b) study in terms of the sensitivity of the test. The study showed semantic judgement secondary task to be more sensitive when performance was compared between quiet and noisy conditions. In addition, semantic judgement task was shown to be sensitive for SNR changes compared to complex visual probe task. As the SNR became poorer the semantic judgment task became more sensitive compared to complex visual probe. The authors again state that the increased depth

of processing of semantic judgement due to linguistic processing as a reason for the better sensitivity of the semantic judgement task.

In another study conducted by Wu et al. (2014) two secondary tasks were used for the same subjects. One was driving simulation and the other was visual task. The results of the study showed both the tasks as sensitive to measure listening effort. Thus, as of now it is not clear what type of secondary task is more suitable and more sensitive to measure listening effort across different task demand conditions (Gagné et al., 2017).

ii. Outcome measures: Both the accuracy of secondary task performance and reaction time measures are used as indicator of listening effort. Studies have shown both measures to have similar pattern of response. Studies by Gosselin and Gagne, (2011a, 2011b) employed tactile pattern recognition secondary task to study the effect of age and mode of stimulus presentation (auditory vs. audio-visual). They used both accuracy of tactile pattern recognition and response time as outcome measures. The results showed both outcome measures to have similar trend and both outcome measures showed increment in listening effort in older age group compared to younger age group participants (Gagné et al., 2017).

iii. Metric of measurement: The dual-task response cost is considered as the indicator of listening effort. Increase in the dual task cost represents increase in listening effort and decrease represents decrease in listening effort. However, the magnitude of change in outcome measure needs to be interpreted with reference to single task baseline. For example, a dual task cost of 10ms (reaction time RT) can have different interpretation if the baseline value is 50ms (RT) (that is 20% change from

baseline) versus when baseline value is 200ms (RT) (that is 5% change from baseline). Hence, it is recommended to use proportion of dual-task cost (pDTC) instead of raw values. Studies conducted by Gosselin and Gagne (2011a, 2011b) have used pDTC to interpret the results, where pDTC is the ratio of dual task cost to the baseline value.

Furthermore, in some instances during dual-task paradigm the performance on primary task changes along with secondary task performance across different test conditions, especially when performance is compared between different SNRs. Here either the dual-task cost of primary or secondary task or both can be used to show changes in listening effort (Gagné et al., 2017). Gagne et al. (2017) propose the use of combined dual cost that is addition of pDTC of primary task and pDTC of secondary task as another option of representing data.

Important factors to consider while measuring listening effort with dual-task paradigm

From literature it can be inferred that there are wide variations in the way the dual-task paradigm has been employed to measure listening effort. The type of secondary task used, the materials used for primary task, the concurrent or sequential response delivery are some examples of variations. Also, there is no clear evidence as to which method and material is superior, suitable and more sensitive across different independent variables. However, it is not appropriate to assume that these factors have no influence on the results of the study (Gagné et al., 2017). Hence, it is important to choose methods based on the purpose and needs of the study. The following are few methodological factors which need consideration to better design a study.

a. Material selection: The factors to be considered while selecting the material is the age and vocabulary knowledge of the population. If the population is children, it is important to understand the auditory experience and vocabulary of the group. If the experience is less and the participants have restricted vocabulary, then high probability word recognition can be a better choice instead of sentence recognition.

b. SNR: SNR selection should be based on the purpose of the study. If the authors intend to test subjects across SNRs then set (constant) signal-to-noise levels can be used. Otherwise, varying SNR which result in equivalent performance across subjects can be used. There is no evidence in the literature to show which method is better or sensitive to measure changes in listening effort. However, if the study design allows then it is preferable to use both methods to understand how listening effort changes with changing speech recognition performance and with equivalent performance across subjects (Gosselin & Gagne, 2011a, 2011b). Using both methods helps to substantiate the results obtained in the study (Gagné et al., 2017). Desjardins and Doherty,(2014) used dual task paradigm to examine the benefit of SMNR in old hearing-impaired individuals. They used visual motor tracking method as secondary task to estimate changes in listening effort across conditions. The percent of time the mouse was on the target was considered as the outcome measure. The results of the study showed improvement in listening effort when the speech recognition scores were near 50% with SMNR compared to no SMNR. However, there was no improvement in listening effort when the speech recognition performance was around 78%.

That is there was no significant difference in performance with and without SMNR. One of the possible reasons for the findings could be that the SNR 78% had poor sensitivity compared to SNR 50% to listening effort change due to ceiling effect. Thus, it is important to select an SNR which helps to avoid ceiling effect. Similarly, study conducted by Wu et al. (2014) reported a decrement in listening effort (or decrease in reaction time) at the poorest SNR used in the study. This observation could be because of the interaction of motivation with speech recognition task. When the speech recognition task becomes too difficult there is possibility that the subject loses motivation to participate in the primary task resulting in improvements in the secondary task performance (floor effect). Thus, it is again important to choose SNR which will avoid floor effect.

c. Linguistic context: Older individuals with poor cognitive skills might find it difficult to perform speech recognition task with low-probability stimuli compared to high probability stimuli. [Note: High probability stimuli are those which are frequently encountered words or sentences and loaded with semantic cue compared to low probability stimuli]. This might prevent to achieve the performance criteria set for the primary task (if it is equivalent performance method). Thus, the researchers may consider using both material or the better of the two in case of persons with cognitive impairment is considered as study population.

d. Secondary task type: The selection of secondary task depends again on the age range of the study population. Pattern recognition task may not be appropriate

for children who are younger if they are still in the pattern recognition developmental stage. Similarly, if older population is considered their dexterity, visual acuity and tactile sensitivity can affect the response as secondary task may require persons to involve in motor activity (visual motor tracking) or engage in visual/tactile tasks. Thus, it becomes necessary consider which task is more appropriate for the population or which factors (motor skills, visual acuity, tactile perception) need to be kept uniform across participants, as this may introduce random noise or high variance in the data.

When the sensitivity of the test is considered, two studies show secondary task requiring semantic judgement to be more sensitive to SNR changes (Picou & Ricketts, 2014a; Strand et al. 2018). However, owing to the wide variation in the use of secondary task further research is needed to understand the role of secondary task which require linguistic processing and auditory processing on the sensitivity of the test.

e. Outcome measure: Accuracy and response time are the two outcome measures used. The response time is a more reliable measure if closed-set speech recognition test is used as a primary task. Because, in closed set speech recognition the person will provide response in the form key press or touch and this can be considered as a reference point to calculate response time (Gagné et al., 2017). Further, if the study population is young subjects or elderly individuals, then using both outcome measures can be useful. Because, if participants are not able follow the instruction of ‘responding as fast as possible’ due to attentional issues the authors will have at least another measure

to rely on. Thus, in such cases secondary tasks where accuracy measurement is possible (for example, semantic judgement, pattern recognition etc.) should be used.

Working memory tests

Working memory tests are behavioral measures used to assess listening effort. Working memory is considered as a factor which can predict the speech recognition scores in difficult listening situation. Working memory is correlated to speech recognition scores in the presence of noise (Kraus, Strait, & Parbery-Clark, 2012). The working memory is necessary for the processing and storing speech information. In literature researchers have used multiple working memory tests to measure listening effort. It ranges from simple recall tests to procedures that require updating and inhibition processes to engage in the test. The n-back digit span test, forward digit span test, backward digit span tests are the simple recall measures where subject is required to repeat the number presented to them through the auditory modality. However, these tests are found to be less sensitive to measure changes in listening effort across different conditions (Strand et al., 2018). Thus, more complex working memory tests such as listening span test, cognitive spare capacity test were used to measure listening effort.

In listening span test the participant is asked to recognize the last word of the sentence and then recall those words after they have heard a certain number of sentences (or block of sentence). When the task difficulty for speech recognition increases due to poor signal to noise ratio or due to low predictability of the sentence the cognitive resources are utilized more for recognition of word leaving little resources for storing the

word affecting recall scores (Johnson, Xu, Cox, & Pendergraft, 2015; Pichora-Fuller et al., 1995; Smith, Pichora-fuller, & Alexander, 2016). The reduction in recall score with increase in task load is considered as an indication of listening effort.

Pichora-Fuller, Schneider, and Daneman (1995) investigated last word recall task in older and younger adults using SPIN-R sentences at different SNRs (0, +5, +8, only speech). The older adults remembered fewer words compared with younger adults and addition of noise reduced recall scores in both young and older adults. Based on the results, the authors concluded that the age-related compromised upstream processing of auditory information and the background noise affects central processes such as storage and retrieval functions of working memory.

The Word Auditory Recognition and Recall Measure (WAARM) test was developed by Smith, Pichora-Fuller and Alexander (2016) to increase the sensitivity of traditional word recognition test. The study introduced alphabet judgement task in addition to recall task and found more recall cost (reduction in word recall scores) with addition of alphabet judgement task. This again shows that the reduction in recall scores is an indication of mental effort due to unfavorable allocation of cognitive resources to recall task.

Updating and inhibition are two cognitive processes which interact with the memory capacity. The Cognitive Spare Capacity Test (CSCT) includes updating and inhibition processes along with recall task. In CSCT test the person will be presented with digits spoken by a female and a male. The subject will be asked to remember the odd/even word spoken by either male or female. This involves inhibition process as the

subject must ignore all other words other than the requested word. To involve updating process the subject may be asked to recall the last word and odd/even word spoken by the female or male voice. Here again the number words spoken by male or female voice needs to be varied. Including the updating the process along with inhibition process increases the complexity of the task.

A study conducted by Strand et al. (2018) used the recall measure with updating process (Running Memory Test), listening span task and CSCT to measure listening effort in normal hearing condition across different speech conditions (speech perception in quiet, speech perception in noise). The aim of the study was to find the convergent validity of different listening effort measures (behavioral, physiological and subjective report). The results revealed that all of the working memory tasks were sensitive to SNR changes. Among these tests running memory test had more effect size compared to listening span test and CSCT. This finding was against the assumption that more complex task would be more sensitive to changes in task demands, because, the running memory task was relatively simple compared to listening span task and CSCT. Thus, the authors say the longer words used for the running memory test and less predictability of the words used for the test as a potential reason for the test being more sensitive. The results also revealed a good correlation among working memory tasks.

Working memory test has been shown to be sensitive to changes in task load and internal factors such as SNR, context (perception of low and high probability sentences), and age as a listening effort measure. The major advantage of working memory test is that it can be easily adapted in clinical set-up. The traditional speech audiometry consists of speech recognition task which can be easily modified to involve recall process. Also,

the interpretation of the results with working memory test is easier and data can be analyzed along with test administration unlike dual-task method which involves complex data analysis procedure. However, currently there are no standardized working memory tests available to measure listening effort. Thus, there is a need to develop such test (Strand, et al. 2018).

Pupillometry: Physiology

Pupillometry is considered as an indicator of cognitive processing load (Kramer, Teunissen, & Zekveld, 2016). The pupil constriction is considered as a result of parasympathetic activity. The pupil dilation is associated with either activation of sympathetic nervous system or inhibition of parasympathetic nervous system (Winn, 2016). Thus, pupil response is a combined entity of sympathetic and parasympathetic nervous system activity. A study in monkeys has shown activation of noradrenergic fibers of coeruleus nucleus to correlate with pupil dilation and effort related energizing activity. Thus, it is believed that pupil dilation is a result of activity in the coeruleus nucleus of sympathetic nervous system. As this is a response to the activation of the autonomic nervous system it is a physiologic response to arousal or stress (Strand et al., 2018)

The different parameters of the pupillary response are believed to represent different activity. For example, the peak pupillary diameter is assumed to represent the momentary load and the resting state pupillary diameter, (before and after the stimulus presentation) is assumed to represent the resting activity. The maximum pupil dilation is

around 0.6mm which is reported to occur 500 ms to 2000 ms post stimulus onset (Winn et al., 2015).

Benefits of pupillometry

Pupillometry is one of the objective measures of listening effort. This is one physiological measure that is shown to be more consistent in measuring listening effort across different conditions. According to FUEL model, the pupillometry can be used to measure the effect of task demand and also the effect of motivation on listening effort. The studies have used pupillometry to measure the effect of task demands. The benefits of pupillometry can be listed as follows,

Multiple applications of pupillometry

a. *Sensitive to task difficulty (different SNR conditions):* A pilot study by Kramer et al. (2016), examined the effect of different signal-to-noise ratios on listening effort as measured by pupillometry. The results showed that the persons to have smaller pupil diameter in difficult SNR condition compared to better SNR condition. The possible reason for the finding as mentioned by the authors was the tendency to quit in difficult situations. Thus, the study though had only ten subjects (normal hearing) showed pupillometry to be sensitive to stimulus related task demand. Similarly, other studies conducted by Zekveld and colleagues (2010, 2014) also has shown pupillometry to be sensitive to SNR changes (Zekveld & Kramer, 2014; Zekveld et al., 2010). The results of these studies show a non-linear relationship between task load (SNR conditions) and pupil diameter change.

b. *Sensitive to spectral degradation:* Another study conducted by Winn et al. (2015) examined the response change rate of speech intelligibility and listening effort (pupil diameter change) with increase in the number of electrodes in vocoded speech. With increase in electrode number the speech intelligibility score increased up to certain level. However, listening effort improved beyond the level reached by speech intelligibility indicating pupillometry to be sensitive to spectral degradation more than speech intelligibility. In addition, as mismatch between electrodes to place mapping of frequency is considered a reason for poor spectral resolution in persons with cochlear implant, there is a scope in utilizing pupillometry for finding the better frequency allocation during programming. However, further research is required to confirm this hypothesis.

c. *Sensitive to contextual load:* Winn (2016) investigated the phenomenon of release from processing load when there is contextual cue in normal hearing individuals and persons with hearing loss (who are using cochlear implants) in unprocessed condition (original sentences) and degraded signal. The results of pupillometry showed release from processing load when there were contextual cues in both normal hearing individuals and cochlear implant users in unprocessed condition. The reason for cochlear implant subjects to not experience release from processing load in degraded condition was attributed to their ability to not derive the cues in that condition. Similarly, a study by Wingfield has shown pupillometry to be sensitive to syntactic complexity.

d. *Sensitive to spatial separation and types of noise:* Studies by Koelewijn and colleagues (2012, 2015) has shown pupillometry to be sensitive to spatial

separation of signals and single talker noise compared to continuous noise or speech shaped noise (Koelewijn, de Kluiver, Shinn-Cunningham, Zekveld, & Kramer, 2015; Koelewijn, Zekveld, Festen, & Kramer, 2012). Though speech perception was better with single-talker noise (as persons can make use of the gaps in the noise to get necessary cues), pupillometry showed higher listening effort for single-talker noise. Based on the above argument, it can be hypothesized that listening effort measured by pupillometry can be sensitive to detect the effect of fundamental frequency and harmonicity in stream segregation and their influence on processing load. Thus, in summary, as pupillometry is shown to be sensitive to depict processing load or cognitive load in various subject related, task demand related and contextual factors it can be a reliable objective measure of listening effort in both children and adults.

Task related factors

e. **No interference from subjectivity:** Pupillometry is an objective method of assessing listening effort. Unlike dual task paradigm pupillometry is not affected by the multi-tasking ability of the person. In dual-task if a person has problem in engaging multi-tasking this may affect the results. But this drawback is not there for pupillometry as it's a single task test.
f. **Miscellaneous:** Other behavioral measures like working memory (reading/listening span tests etc.,), dual-task paradigm etc., will be affected by the subjectivity of the examiner such as the way examiner perceives the verbal responses of the persons etc., These difficulties are not seen for pupillometry as analysis majorly objective.

Challenges and solutions

Although there are multiple advantages of pupillometry it also has its own limitations and challenges which makes its adaptation difficult in clinics in the present time. The challenges that are faced with pupillometry and possible solutions for the same can be listed as follows,

a. **Off-line analysis of response:** The major challenge in utilizing pupillometry in clinical set up is the analysis method used. Currently off-line analysis methods are being used in the research studies. However, on-line response analysis and immediate disclosure of results is the prime requirement of clinical setting. Hence, off-line analysis is a major drawback in adopting pupillometry for clinical practice.

   Solution: The possible solution for this problem will be standardization of analysis procedure used and development of pupillometry devices for the sole purpose of clinical use with on-line analysis methods.

b. **Influence of subjective factors:** Though pupillometry is resistant to drawbacks of subjective analysis methods, the pupil response is influenced by certain subjective factors. With increase in age the pupil dilation reduces (Winn et al., 2015). This makes it hard to compare the pupil responses across different age groups. The pupil response shown to vary through the hormonal cycle.

   Solution: A researcher needs to consider these factors while conducting research and while interpreting results.

c. **Effect of material and test procedure:** The pupillometry responses are influenced by the affective processing. This make the responses to be

susceptible to the material used or to the stress that procedure creates. In addition, Winn (2016) report the length of the test to affect the pupil responses as it creates fatigue and may result in less arousal.

Solution: Again, the researcher will need to keep these factors in mind while administering the test and while comparing the results with the results of other tests. Winn (2016) suggest using short stimulus lists for pupillometry to avoid any negative effects on the responses.

d. **Effect of light:** The effect of light on pupillary responses dominates the cognitive load. If luminance is not taken care of then it can result in floor and ceiling effects where observing the small changes in pupil dilation becomes difficult.

   Solution: To avoid the effect of luminance the color of the screen can be changed from black to white gradually to find a median position of pupil dilation (Winn et al., 2015).

e. **Lack of consistency in analysis methods:** Currently there are no standardized methods of data analysis while analyzing pupillometry data. Majority of the studies use peak pupil dilation, average pupil dilation as parameters to investigate the effect of independent variables. However, the criteria to select peak pupil dilation such as the window size used for picking the peak dilation, the criteria used to average diameter are all different. Even artifact recognition criteria, artifact removal criteria are all different across studies. The study conducted by Strand et al. (2018) did not use artifact rejection. Whereas studies conducted by Picou and Ricketts (2014a), Steel et al. (2015), Winn (2016) etc.,

use different methods of artifact rejection. Thus, until these procedures are standardized there will be problems of reliability. The sensitivity of this procedure may also vary because these reasons.

Solution: There is a need to standardize the analysis procedure used. Also, there is a need for research to understand the effect of different analysis criterion used on the results of the test.

f. **Subject or data attrition:** In studies where pupillometry is involved the data collection is not possible because of various reasons. The subjects may not be able to follow the instructions, the data attrition due to technical reasons, artifacts (eye blinking). In a study conducted by Strand et al. (2018) involving 111 subjects nearly 10% of subjects' data was not used for analysis because of the above-mentioned reasons. Similarly, due to artifact rejection there arises a need to remove nearly 20% or even more of the data collected for a subject (Winn, 2016).

   Solution: In research concurrent data analysis can be a solution to avoid the problems of data attrition. If analysis shows significant data attrition the researcher will have the option of collecting more data. However, this remains a challenge in clinical population as if the procedural difficulties preclude data collection in a person the use of pupillometry will have to be replaced with other feasible objective measures of listening effort.

Subjective methods

Subjective self-report methods rely on the direct expression of subject's experience of listening task. The self-report methods have the advantage of no

technology requirement and good face validity (Seeman & Sims, 2015). These are either single measures or questionnaires assessing the listening focused questions about effort in daily life (about daily activities). These questions are usually rated using rating scale (with varying ranges and divisions), where zero represents no effort to the maximum scale point represents highest effort perceived. Self-report measures include procedures where participants either provide an immediate feedback about the amount of effort perceived during an activity or a retrospective perception of the listening experience. The commonly used scales are as follows,

i. Speech, Spatial and Qualities questionnaire (SSQ): The qualities sub-section of the SSQ questionnaire has three questions which is regarding the listening effort. These questions are commonly used to assess the subjective ratings of listening effort. These questions are rated on a ten-point scale where zero represents no effort and ten represents maximum listening effort.
ii. NASA-TLX: This is effort measurement scale used in the studies (Strand, et al. 2018). This is a visual rating scale with 27 divisions without numerical marking and subjects will be asked to mark a point on the scale which corresponds to their perception of effort. NASA-TLX can be used as task specific subjective rating scale and can be used to measure listening effort perceived during a laboratory speech perception task.
iii. Single dimensional questions (single questions) are the questions that require patients to rate the amount of effort experienced following a speech perception task (usually a laboratory-based speech perception task). There are multiple variations of single dimension questions. For example, (1) did you perceive effort

while listening to speech in noise, (2) how do you rate the ease of listening experienced during the task, (3) how much mental work was required to complete this task, etc.

The advantage of the self-report tools is that they are easy to administer, and less time consuming. It requires less raining to teach administration of these tools for personnel working with persons with hearing loss. However, there are no standardized subjective scales available currently.

Purpose of the study

Sensitivity of listening effort measures

Having the information on test efficacy is critical for the clinical adaptation of listening effort measures as part of hearing test battery. There is an abundance of measures that are used to measure listening effort as evidence in the literature review section. The three major classes of measures are physiological measures, behavioral measures and subjective rating measures. Various studies have tested the sensitivity of these measures in examining the effect of internal and external factors. However, it is difficult to compare the effect sizes across studies due to varying populations and methodological differences. A systematic review and meta-analysis by Ohlenforst et al. (2017), analyzed the listening effort literature to examine the evidence available for different listening effort measurement tools. The main purpose of the study was to investigate the evidence available to support the two hypotheses: (1) listening effort in persons with hearing loss is more compared to normal hearing individuals, (2) hearing aid helps to reduce listening effort. In this systematic review the authors provide a

comparison of outcomes obtained with subjective measures and objective measures. The results show a large proportion of objective tests, both behavioral and physiological tests (15 out of 23 or approximately 65%) to show significant improvement with hearing aid treatment compared to subjective tools (17 out of 33 or 51%). However, after conducting the post analysis of the quality of the subjective and objective methods the authors note the wide variation in the methodology used for subjective tools and objective tools and they mention the wide variability in the methodology as the main reason to not able to compare studies and build evidence for listening effort test.

Few studies examined the comparative sensitivity of listening effort measures (Alhanbali, Dawes, Millman, & Munro, 2019; Johnson et al., 2015; Seeman & Sims, 2015). Alhanbali and colleagues (2019) simultaneously measured pupil size, electroencephalographic alpha power, skin conductance, and self-reported measure of effort in 116 participants with normal to severe hearing loss. The testing was conducted at SNR corresponding to 71% performance on digit recall task and results showed pupillometry to explain higher percent of variance compared to alpha power changes and subjective rating. Study by Johnson et al. (2015) showed subjective rating scale to be more sensitive in reflecting changes in SNR than listening span test in normal hearing individuals (N=30). Seeman and Sims (2015) compared physiological measures (skin conductance, hear rate, and heart-rate variability), dual-task measure and subjective ratings at two SNRs (+5 and +15 dB) in normal hearing individuals and found subjective rating compared with physiological or dual-task measure to be more sensitive.

It is difficult to select tools for clinical use because: the studies examine tools efficiency at a small range of task difficulty (Alhanbali et al., 2019; Seeman & Sims,

2015); the studies compare tools with selective underlying constructs (Johnson et al., 2015). Comparing the physiological, behavioral and subjective measures is important as they have different strengths and weaknesses. The physiological measures provide temporal precision and effort change across time, but these measures require dedicated equipment. The behavioral measures represent real life experience, but internal factors like ability to perform multiple tasks, baseline working memory capacity. The subjective measures give face validity as it measures the experience of persons but may get influenced by the subjective bias and misperception of the questions.

The present study aims to examine the efficiency of pupillometry, working memory, and subjective rating scales in demonstrating the effect of signal to noise ratio and presentation levels. The pupillometry was selected as it has shown consistent pattern of results for the effect of SNR. The subjective rating scale was selected as it is most easy to administer, cost effective and time efficient measure. The working memory test was selected as it is an easy behavioral measure to incorporate along with already existing sentence perception task in the clinics. Also, these three measures were selected as they represent different classes of listening effort measurement methods.

Objective, behavioral and subjective measures: Underlying construct

A recent study by Alhanbali, Dawes, Millman, and Munro (2019) explored the underlying constructs of pupillometry, electroencephalography, skin conductance, and subjective rating of effort. They simultaneously measured listening effort using all four measures in 116 individuals with normal to severe hearing loss. The weak correlation between the measures despite good reliability of the measures was considered as an evidence of multi-dimensionality of listening effort measures. The authors conclude by

saying the different measures of listening effort should not be used interchangeably as they have different constructs.

Furthermore, one of the frequent observations is the discrepancy in objective and subjective methods results in measuring listening effort (Desjardins, 2016; Desjardins & Doherty, 2014; Ohlenforst et al., 2017; Pals et al., 2013; Pichora-Fuller et al., 2016; Strand et al., 2018). This discrepancy is noted in both cochlear implant and hearing aid literature. A study conducted by Pals et al. (2013) showed the discrepancy between subjective and behavioral measurement method in cochlear implant individuals. The authors investigated the effect of spectral degradation on speech recognition, listening effort and subjective rating. The subjects had normal hearing sensitivity and they were presented with vocoded speech with varying number of channel information. The hypothesis was with increase in the number of channels in the vocoded speech there will be increment in speech perception, listening effort and self-perceived effort. The results showed speech recognition scores and subjective perception to improve till six channels; however, the listening effort as measured using dual-task paradigm improved beyond six channels and up to eight channels. This shows that dual-task paradigm is more sensitive to the effects of spectral degradation compared to subjective ratings.

Similarly, studies conducted by Desjardins and colleagues (2014, 2016) showed discrepancy between subjective and objective methods in hearing aid users. The study conducted by Desjardins and Doherty (2014) examined the effects of digital noise reduction on speech recognition, listening effort and subjective rating and ease of listening in persons with hearing impairment. The authors used dual-task paradigm to measure listening effort with visual motor tracking as the secondary task. The results

showed significant improvement in visual motor tracking performance when digital noise reduction algorithm was on, showing improvement in listening effort only in difficult speech recognition condition. However, the self-report measure used (question regarding ease of listening) did not show significant improvement.

In general, the subjective measures are shown to have better sensitivity in measuring listening effort changes across different SNRs and other task demands (Strand et al., 2018). This observation is attributed to the good correlation between subjective measure and speech recognition scores. The good correlation between subjective rating of listening effort and speech recognition scores show that the persons' rating is influenced or biased by their perception of their performance level (Moore & Picou, 2018). Though in hearing aid and cochlear implant literature subjective methods do not emerge as sensitive measure compared to objective method this observation can be an explanation for the discrepancy found between subjective and objective methods. The results of the Pals et al. (2013) study can be explained by this observation. In Pals et al. (2013) study both speech recognition scores and subjective rating scores saturates by six channels and this result could be because that the subjects are judging their listening effort based on their performance in speech recognition test.

On the other hand, behavioral measures and objective measures are shown to have good correlation. Both dual task paradigm (reaction time measurement) and Pupil dilation have shown non-linear relationship across SNRs. The study by Wendt et al. (2018) and Wu et al. (2014) have shown maximum effort in the region where sentence perception score is between 30% and 50% or 70% and decreased effort beyond this range. Hence, it is still not clear as to which method or combination of measures will better represent the

listening effort during speech perception. Also, as general mental status and other internal factors can affect these measures it is important to simultaneously administer these measures. In the current study, the sentence perception task, working memory task and pupillometry are measured simultaneously to control the internal factors.

Reliability of measures

Alhanbali et al. (2019) examined the reliability of pupil dilation, electroencephalography, skin resistance and subjective measure listening effort and fatigue. The results showed except for skin conductance and subjective measure fatigue rest all measures to have good reliability (minimum ICC: 0.71). Establishing the reliability of any outcome measure is crucial for its clinical adaptation. As decision regarding rehabilitation is made on a case by case basis and to monitor the benefit of any rehabilitation program overtime, it is necessary for the outcome measures to have good reliability. In addition, to examine the underlying constructs of any measure it is very necessary for those measures to have good reliability as poor reliability can result in weak correlation between different measures of listening effort. In literature, there are very few studies which have examined the reliability of the listening effort measures. Hence in the present study, the reliability measure of pupil dilation, working memory and subjective rating will be estimated by re-administering the test in 20% of the sample size.

Audibility and listening effort

The primary goal of rehabilitation with hearing aids is restoration of audibility. During hearing aid programming, it is important to confirm if the patient have access to speech at a comfortable level. The major validation tools used to check restoration of

audibility are real-ear measurement and speech perception tests. Studies have shown even with same speech perception scores persons may employ different levels of effort to attain the same level of performance. Hence, it may be beneficial to measure listening effort during programming to make sure appropriate amount of gain is provided to ease the communication process. To use listening effort measures to validate hearing aid gain, it is important to establish the relationship between presentation level and effort. There are equivocal results regarding the effect of presentation level on pupil responses. A study by Liao, Kidani, Yoneya, Kashino, and Furukawa, (2016) examined the effect of presentation level of tones, noise-bursts on pupil response and observed louder signals to be associated with larger pupil responses. In contrast, a study by Zekveld et al. (2010) the baseline level did not vary significantly when pupil responses were measured at different noise levels while keeping the sentence level constant. Also, there are no studies which examine the effect overall presentation of speech and noise presentation in a speech perception in noise task. In the present study, the effect of presentation level will be examined on pupil dilation, working memory, and subjective measure.

Research questions

1. How listening effort varies across different SNR conditions as measured by pupillometry, working memory and subjective rating scale?
2. How listening effort varies across presentation levels as measured by pupillometry, working memory and subjective rating scale?
3. Is there any interaction between SNR conditions and presentation levels?
4. Which measure among the three has greater efficacy in reflecting the effect of SNR and presentation level?
5. What is the reliability of listening effort measures?

APPENDIX-2
STIMULUS LAYOUT

Silence															
Noise															
Sentence															
Sentence recognition														Repeat	
Listening span															Recall
Pupillometry				B		Pupil dilation data									
Time (seconds)	1	2	3	4	5	6	7	8	9	10	11	12	13	14- 18	19-33

Each condition started with one-minute pupil calibration and five seconds instruction display on the screen. Each trial began with 5 seconds of silence followed by eight seconds of stimulus. The stimulus had sentence (average 2 seconds in duration) embedded in the middle of eight second noise. At the end of each sentence the participant was given five seconds time to repeat the sentence (Repeat). After five sentences, the participants were given fifteen seconds to recall the five last words from the sentences (Recall). B = Baseline for pupillometry data, Pupil dilation data = The pupillometry data analyzed to obtain peak pupil dilation.

APPENDIX-3

SELECTION CRITERIA QUESTIONNAIRE

Name: …….

Age: ……....

Current Education: …….

Read the questions below and answer by circling the appropriate letter. (Y-Yes, N-No)

1. Are you a native speaker of American English? Y/N
2. Do you speak any other language other than American English? Y/N
   If yes, please mention the languages below …….
3. Are you a trained musician?
   If yes, please mention the duration of training: .....
4. Do you have any difficulty to understand speech in the presence of noise? Y/N
5. Did you have ear infection in past three months? Y/N
6. Are you diagnosed with attention disorder? Y/N
7. Are you taking any medications currently for attention disorder? Y/N
8. Are you diagnosed with Epilepsy? Y/N
9. Are you currently taking medication for epilepsy? Y/N
10. Are you currently under any other medications?
11. Are you sensitive to flickering lights, TV screen or arcade games? Y/N
12. Do you currently have infrared sensitive medical device on your body? Y/N
13. Do you wear glasses or contact lenses to read? Y/N
14. Have you ever had any eye injury or disease?
    If any, please mention the cause.
15. Mention the duration for which you had/have the problem:
    ……

APPENDIX-4

SUBJECTIVE RATING OF LISTENING EFFORT

Participant Code: …….

1. How hard did you have to listen to understand the sentences in noise?

   Low effort High effort

   1 2 3 4 5 6 7 8 9 10

   ● ● ● ● ● ● ● ● ● ●

2. How frustrated, discouraged, irritated, stressed and annoyed did you feel during the task?

   Low frustration High frustration

   1 2 3 4 5 6 7 8 9 10

   ● ● ● ● ● ● ● ● ● ●

3. How often did you give up trying to understand the sentence?

   Never Most of the time

   1 2 3 4 5 6 7 8 9 10

   ● ● ● ● ● ● ● ● ● ●

4. How would you rate your performance on the task?

   Good performance Poor performance

   1 2 3 4 5 6 7 8 9 10

   ● ● ● ● ● ● ● ● ● ●

5. How hard did you have to work to remember/recall the words?

   Low effort High effort

   1 2 3 4 5 6 7 8 9 10

   ● ● ● ● ● ● ● ● ● ●

APPENDIX 5

EXPLORATORY CORRELATION ANALYSIS

Correlation Plot at 6 dB SNR

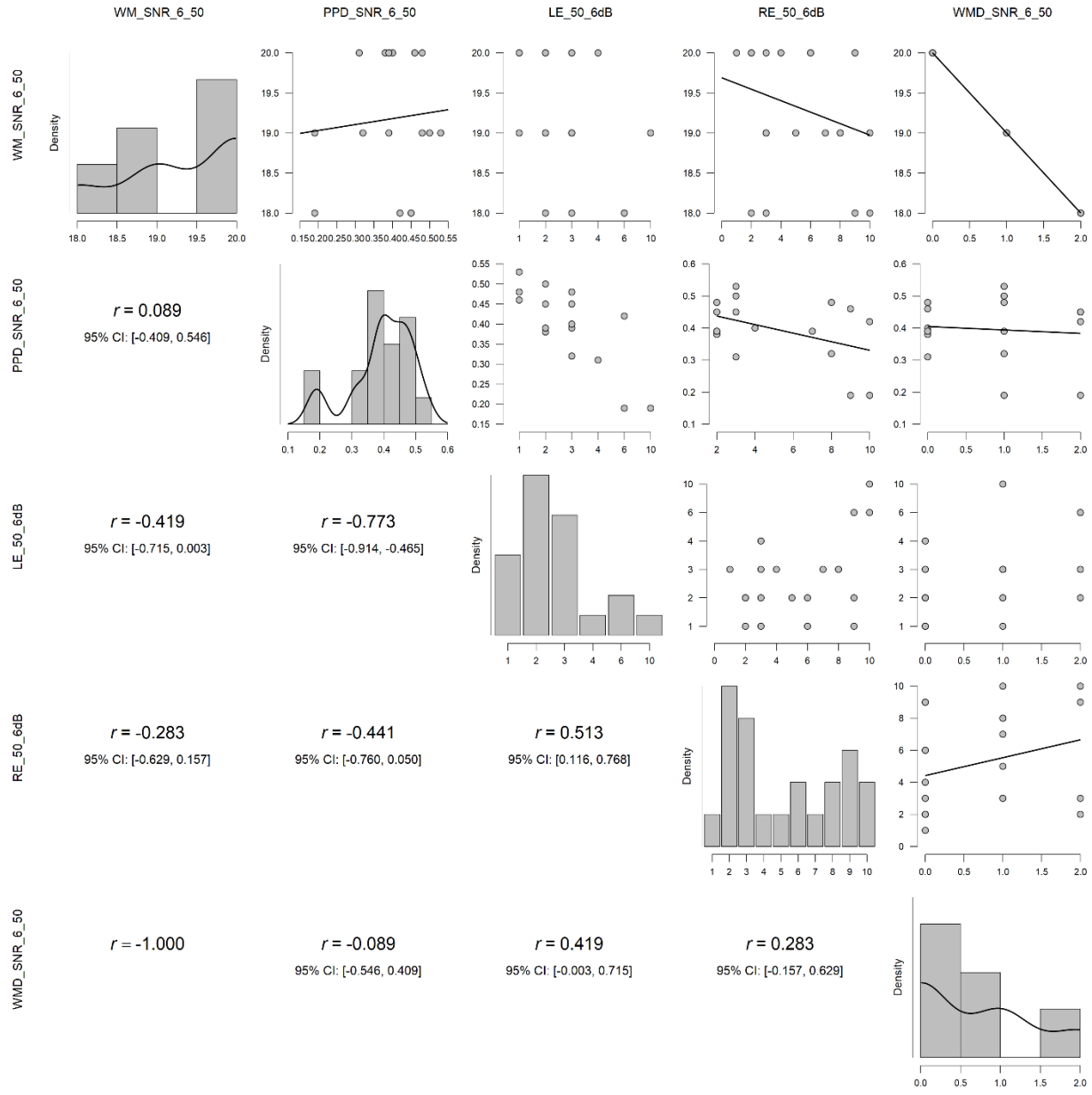

Figure 1: Correlation analysis between listening effort measures at 6 dB SNR

WM= Working memory, WMD= Working memory difference, PPD= Peak pupil dilation change, LE = Subjective rating of listening effort, RE= Subjective rating of recall effort.

Correlation Plot at 3 dB SNR

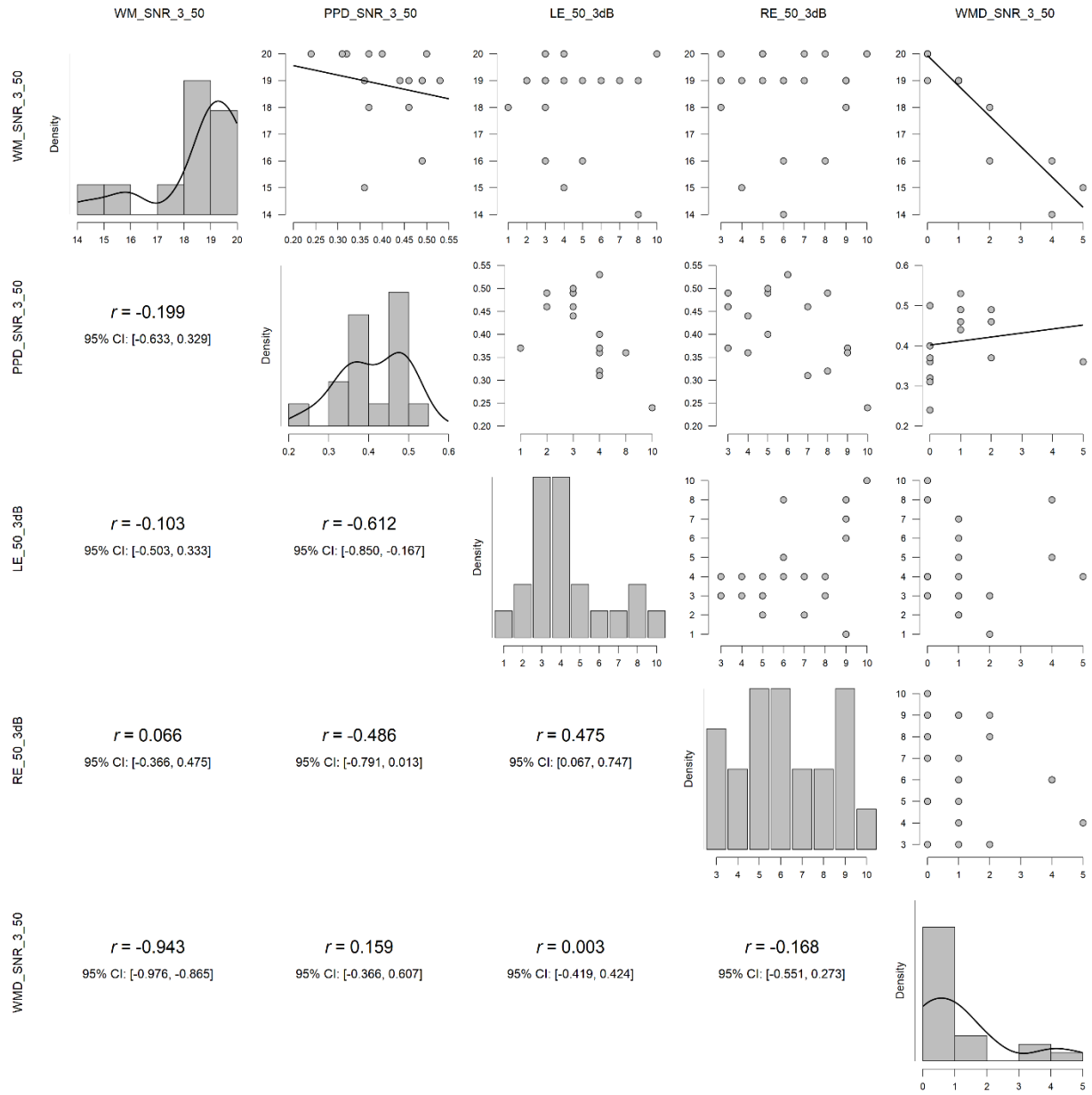

Figure 2: Correlation analysis between listening effort measures at 3 dB SNR

WM= Working memory, WMD= Working memory difference, PPD= Peak pupil dilation change, LE = Subjective rating of listening effort, RE= Subjective rating of recall effort.

Correlation Plot at 0 dB SNR

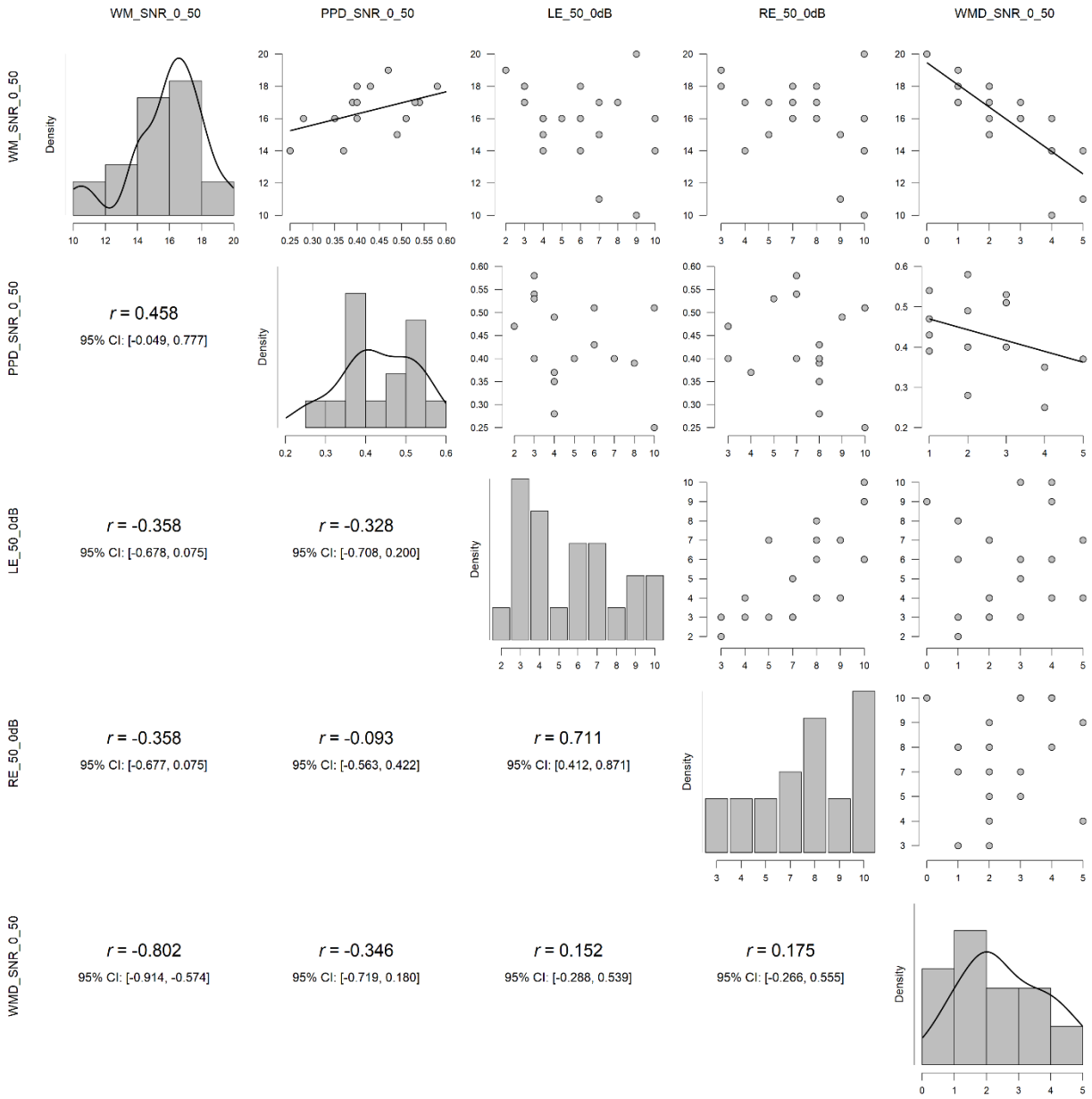

Figure 3: Correlation analysis between listening effort measures at 0 dB SNR

WM= Working memory, WMD= Working memory difference, PPD= Peak pupil dilation change, LE = Subjective rating of listening effort, RE= Subjective rating of recall effort.

Correlation Plot at -3 dB SNR

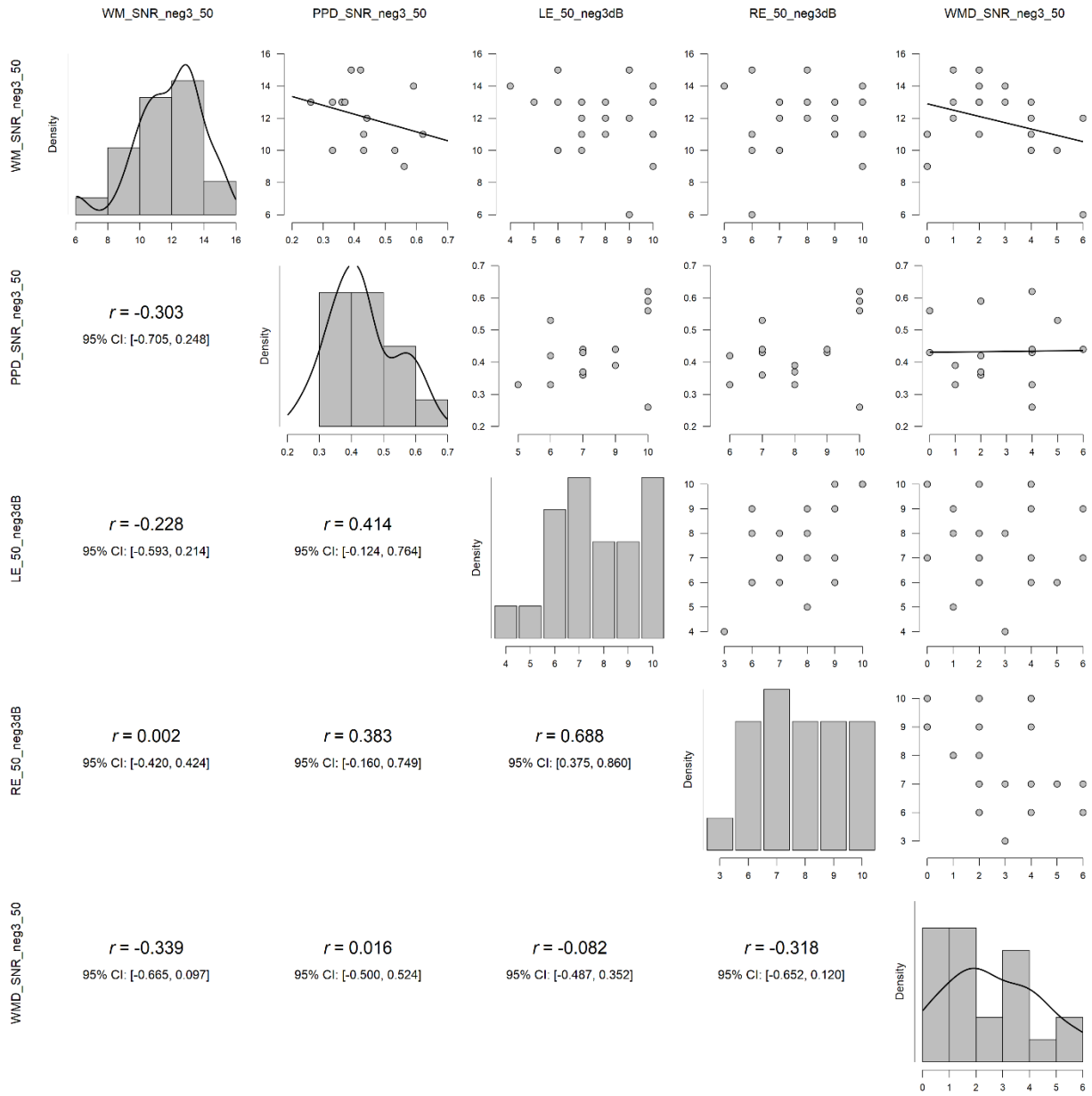

Figure 4: Correlation analysis between listening effort measures at -3 dB SNR

WM= Working memory, WMD= Working memory difference, PPD= Peak pupil dilation change, LE = Subjective rating of listening effort, RE= Subjective rating of recall effort.

Correlation Plot at -6 dB SNR

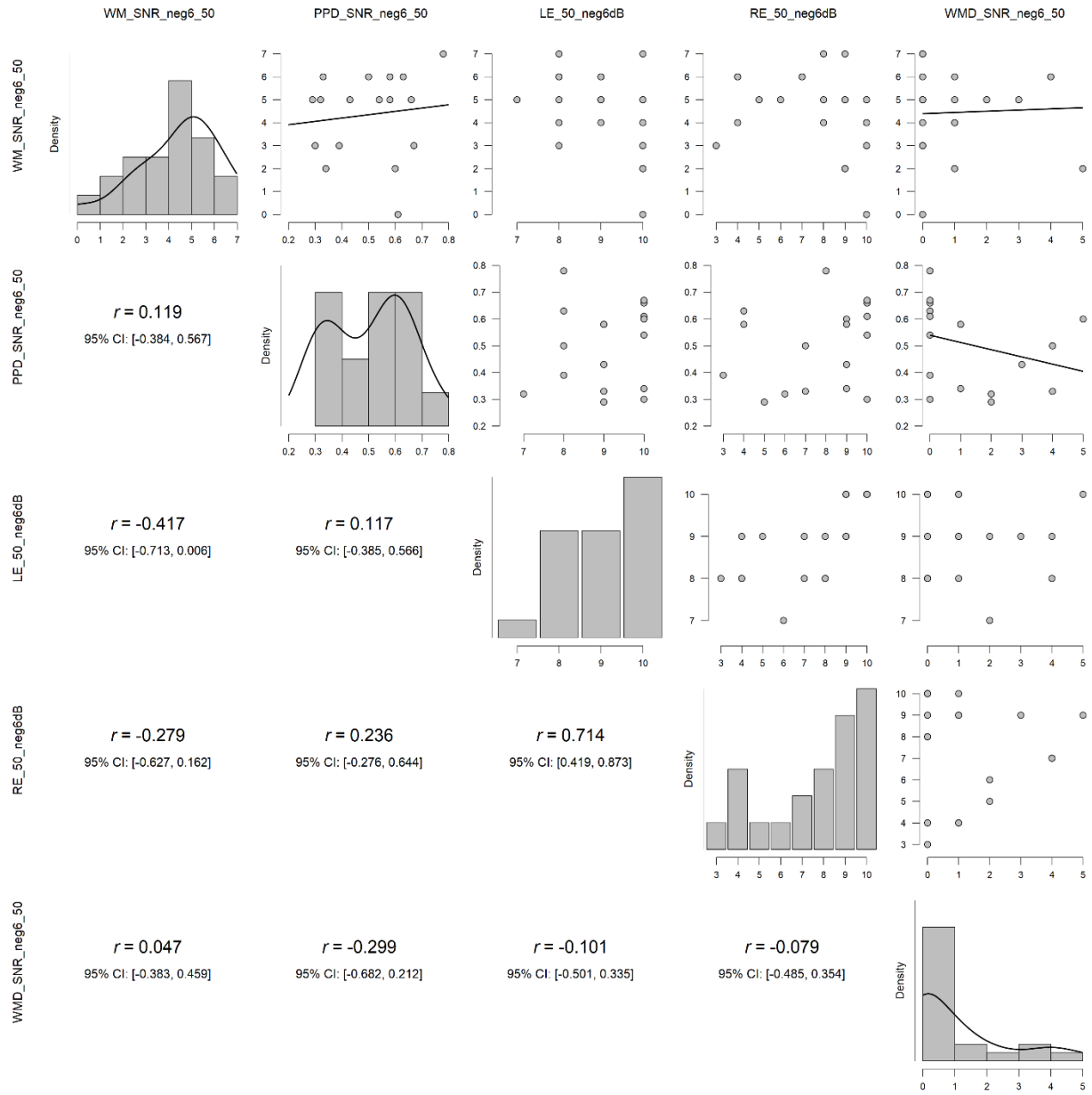

Figure 5: Correlation analysis between listening effort measures at -6 dB SNR

WM= Working memory, WMD= Working memory difference, PPD= Peak pupil dilation change, LE = Subjective rating of listening effort, RE= Subjective rating of recall effort.

Correlation Plot at -10dB SNR

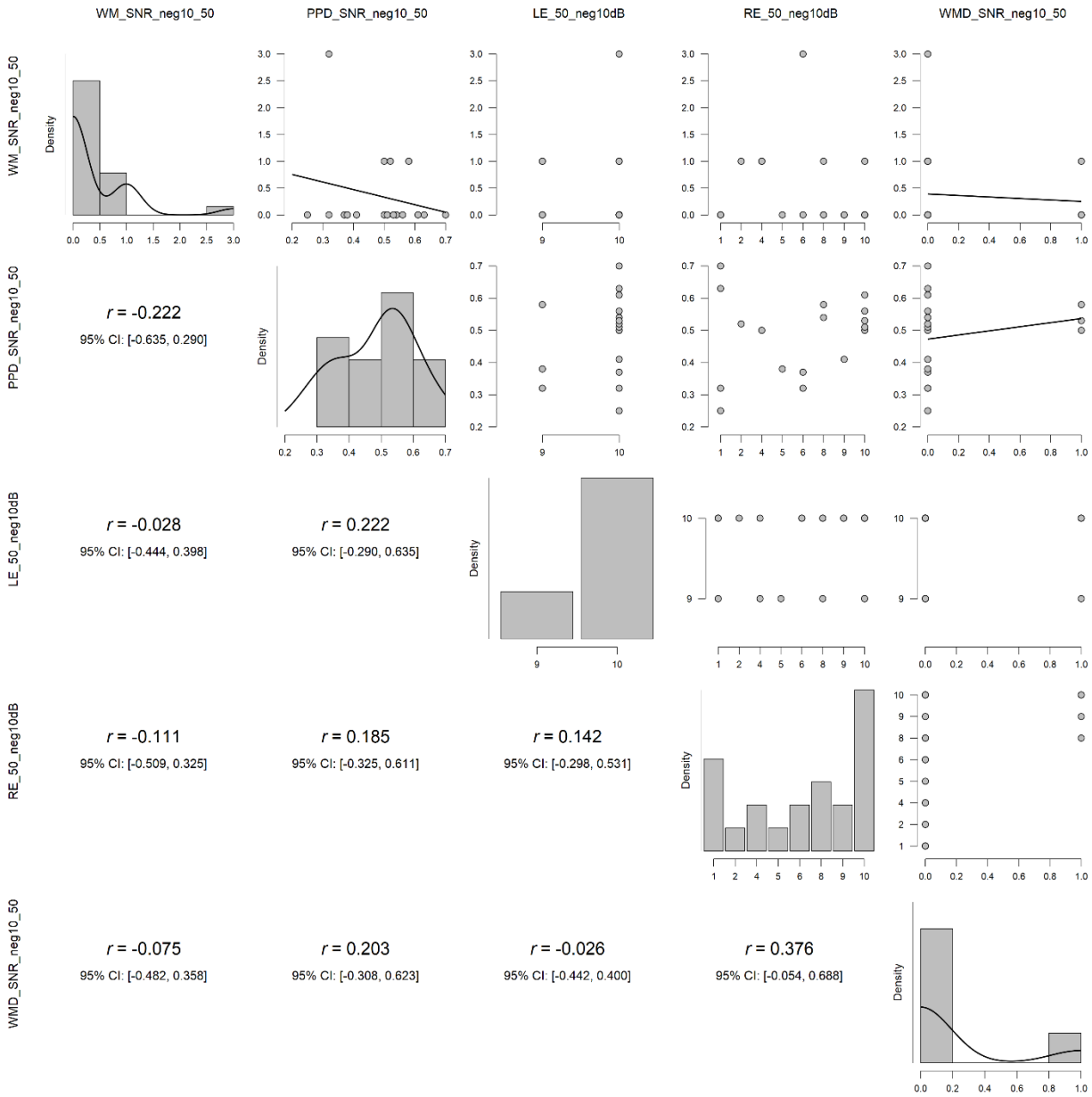

Figure 6: Correlation analysis between listening effort measures at -10 dB SNR

WM= Working memory, WMD= Working memory difference, PPD= Peak pupil dilation change, LE = Subjective rating of listening effort, RE= Subjective rating of recall effort.

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. https://doi.org/10.1097/AUD.0000000000000361

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2018). Hearing handicap and speech recognition correlate with self-reported listening effort and fatigue. *Ear and Hearing*, *39*(3), 470–474. https://doi.org/10.1097/AUD.0000000000000515

Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear and Hearing*, *40*(5), 1084–1097. https://doi.org/10.1097/aud.0000000000000697

Amichetti, N. M., Stanley, R. S., White, A. G., & Wingfield, A. (2013). Monitoring the capacity of working memory: Executive control and effects of listening effort. *Memory and Cognition*, *41*(6), 839–849. https://doi.org/10.3758/s13421-013-0302-0.Monitoring

Antikainen, J., & Niemi, P. (1983). Neuroticism and the pupillary exposure to noise. *Biological Psychology*, *17*, 131–135.

Bala, A. D. S., Whitchurch, E. A., & Takahashi, T. T. (2020). Human auditory detection and discrimination measured with the pupil dilation response. *Journal of the Association for Research in Otolaryngology*, 43–59. https://doi.org/10.1007/s10162-019-00739-x

Bentler, R., Wu, Y. H., Kettel, J., & Hurtig, R. (2008). Digital noise reduction: Outcomes from laboratory and field studies. *International Journal of Audiology*, *47*(8), 447–460. https://doi.org/10.1080/14992020802033091

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear and Hearing*. https://doi.org/10.1097/AUD.0000000000000028

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G * Power 3 : A flexible statistical power analysis program for the social, behavioral, and biomedical sciences, *39*(2), 175–191.

Gosselin, A. P., & Gagné, J. (2011b). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research*, *54*, 944–958. https://doi.org/10.1044/1092-4388(2010/10-0069)a

Gosselin, P. A., & Gagné, J. P. (2011a). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *International Journal of Audiology*, *50*, 786–792. https://doi.org/10.3109/14992027.2011.599870

Guijo, L. M., Horiuti, M. B., & Cardoso, A. C. V. (2019). Measurement of listening

effort using of a dual-task paradigm of Brazilian Portuguese : a pilot study. *CoDAS*, *31*(4), 1–9. https://doi.org/10.1590/2317-1782/20192018181

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society*, 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*(C), 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hornsby, B. W. Y. (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. https://doi.org/10.1097/AUD.0b013e31828003d8

Hornsby, B. W. Y., & Kipp, A. M. (2016). Subjective ratings of fatigue and vigor in adults with hearing loss are driven by perceived hearing difficulties not degree of hearing loss. *Ear and Hearing*, *37*(1), e1–10. doi: 10.1097/AUD.0000000000000203.

Hughes, S. E., Hutchings, H. A., Rapport, F. L., McMahon, C., & Boisvert, I. (2018). Social connectedness and perceived listening effort in adult cochlear implant users: A Grounded Theory to establish content validity for a new patient-reported outcome measure. *Ear & Hearing*, *Ear and Hearing*, *39*(5), 922-934.

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), 1–5. https://doi.org/10.1136/bmjopen-2016-014995

Humes, L. E., Christensen, L. A., Bess, F. H., & Hedley-williams, A. (1997). A comparison of the benefit provided by well-fit linear hearing aids and instruments with automatic reductions of low frequency gain. *Journal of Speech Language Hearing Research*, *40*(3), 666–685.

Johnson, J. A., Xu, J., & Cox, R. M. (2016). Impact of hearing aid technology on outcomes in Daily Life II: Speech understanding and listening effort. *Ear and Hearing*, *37*(5), 529–540. https://doi.org/10.1097/AUD.0000000000000327

Johnson, J., Xu, J., Cox, R., & Pendergrafta, P. (2015). A comparison of two methods for measuring listening effort as part of an audiologic test battery. *American Journal of Audiology*, *24*(September), 419–431. https://doi.org/10.1044/2015

Jonas Brännström, K., Karlsson, E., Waechter, S., & Kastberg, T. (2018). Listening effort: Order effects and core executive functions. *Journal of the American Academy of Audiology*, *29*(8), 734–747. https://doi.org/10.3766/jaaa.17024

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. https://doi.org/10.1080/14992020600754583

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, *51*(3), 1336–1342. https://doi.org/10.3758/s13428-018-1075-y

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2018). Development of an adaptive scaling method for subjective listening effort. *The Journal of the Acoustic Society of America*, *141*(6), 4680-4693. https://doi.org/10.1121/1.4986938

Krueger, M., Schulte, M., Zokoll, M. A., Wagener, K. C., Meis, M., Brand, T., & Holube, I. (2017). Relationship between listening effort and speech intelligibility in noise. *American Journal of Audiology*, *26*(3S), 378–392.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry : A Window to the Preconscious? *Perspectives on Psychological Science*, *7*(1), 18-27. https://doi.org/10.1177/1745691611427305

Liao, H. I., Kidani, S., Yoneya, M., Kashino, M., & Furukawa, S. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychonomic Bulletin and Review*, *23*(2), 412–425. https://doi.org/10.3758/s13423-015-0898-0

Lin, F. R., Niparko, J. K., & Ferrucci, L. (2011). Hearing loss prevelence in the United States. *Archives of Internal Medicine*, *171*(20), 1851–1853. https://doi.org/10.1002/gps.2627.6.

Lunner, T., Rudner, M., Rosenbom, T., Ågren, J., & Ng, E. H. N. (2016). Using speech recall in hearing aid fitting and outcome evaluation under ecological test conditions. *Ear and Hearing*, *37*(Supplement), 145S-154S.

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. https://doi.org/10.3109/14992027.2014.890296

Meister, H., Schreitmüller, S., Grugel, L., Ortmann, M., Beutner, D., Walger, M., & Meister, I. G. (2013). Cognitive resources related to speech recognition with a competing talker in young and older listeners. *Neuroscience*, *232*, 74–82. https://doi.org/10.1016/j.neuroscience.2012.12.006

Moore, T. M., & Picou, E. M. (2018). A potential bias in subjective ratings of mental effort. Journal of Speech, Language, and Hearing Research, 61(9), 2405-2421.

Ng, E. H. N. (2013). *Cognition in Hearing Aid Users : Memory for Everyday Speech* (Doctoral dissertation, Linköping University Electronic Press).

NIDCD (2016). *Quick statistics in hearing*. National Institute of Health. https://www.nidcd.nih.gov/health/statistics/quick-statistics-

hearing#:~:text=About%202%20percent%20of%20adults,older%20have%20disabling%20hearing%20loss.

Nike (2017). Speech in noise mixing, signal to noise ratio (https://www.mathworks.com/matlabcentral/fileexchange/37842-speech-in-noise-mixing-signal-to-noise-ratio), MATLAB Central File Exchange. Retrieved August 1, 2017.

Nilsson, M., Soli, S. D., Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *95*(2), 1085–1099. https://doi.org/10.1121/1.408469

Nunnally, J. U. M. C., Knott, P. D., & Duchnowski, A. (1967). Pupillary response as a general measure. *Perception & Psychophysics*, *2*, 149–155.

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., … Kramer, S. E. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79. https://doi.org/10.1016/j.heares.2017.05.012

Padilla, L. M. K., Castro, S. C., Quinan, P. S., Ruginski, I. T., & Creem-regehr, S. H. (2020). Toward objective evaluation of working memory in visualizations : a case study using pupillometry and a dual-task paradigm. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 332–342.

Pals, C., Sarampalis, A., & Baskent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech Language and Hearing Research*, *56*(4), 1075-1084. https://doi.org/10.1044/1092-4388(2012/12-0074)

Peelle, J. E. (2018). Listening effort : How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204–214.

Peysakhovich, V., Vachon, F., & Dehais, F. (2017). The impact of luminance on tonic and phasic pupillary responses to sustained cognitive load. *International Journal of Psychophysiology*, *112*, 40–45. https://doi.org/10.1016/j.ijpsycho.2016.12.003

Pichora-Fuller, M. K. (2010). Using the brain when the ears are challenged helps healthy older listeners compensate and preserve communication function. In *Hearing Care for Adults: The Challenge of Aging. Proceedings of the Second International Conference* (pp. 53-65).

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., … Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*(Supplement 1), 5S-27S. https://doi.org/10.1097/AUD.0000000000000312

Pichora-Fuller, M. K., Mick, P., & Reed, M. (2015). Hearing, cognition, and healthy aging: social and public health implications of the links between age-related declines in hearing and cognition. *Seminars in Hearing*, *36*(3), 122-139.

https://doi.org/10.1055/s-0035-1555116

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *Journal of the Acoustical Society of America*, *97*(1), 593–608. https://doi.org/10.1121/1.412282

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*, 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model : theoretical, empirical, and clinical advances overview understanding. *Frontiers in Sytems Neuroscience*, *7*(July), 1–17. https://doi.org/10.3389/fnsys.2013.00031

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*(sup2), S99–S105. https://doi.org/10.1080/14992020802301167

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. https://doi.org/10.3766/jaaa.23.7.7

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: effects of background noise and noise reduction. *Journal of Speech Language and Hearing Research*, *52*(5), 1230. https://doi.org/10.1044/1092-4388(2009/08-0111)

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech Language and Hearing Research*, *58*, 1781–1792. https://doi.org/10.1044/2015_JSLHR-H-14-0180

Shehorn, J., Marrone, N., & Muller, T. (2018). Speech perception in noise and listening effort of older adults with nonlinear frequency compression hearing aids. *Ear and Hearing*, *39*(2), 215–225. https://doi.org/10.1097/AUD.0000000000000481

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech Language and Hearing Research*, *61*(6), 1463-1486. https://doi.org/10.1044/2018_JSLHR-H-17-0257

Tabachnick, B. G., & Fidell, L. S. (2013). *Using Multivariate Statistics* (6th Editio). Boston: Pearson.

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research*, *369*, 67–78. https://doi.org/10.1016/j.heares.2018.05.006

Winn, M. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing*, *20*, 1–17. https://doi.org/10.1177/2331216516669723

Wu, Y., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort, *37*(6), 660–670. https://doi.org/10.1097/AUD.0000000000000335.Psychometric

Yeh, Y., & Wicken, C. D. (1984). The Dissociation of Subjective Measures of Mental Workload and Performance. 1-138.

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. https://doi.org/10.1111/psyp.12151

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. https://doi.org/10.1097/AUD.0b013e3181d4f251