**Older adults show a more sustained pattern of effortful listening than young adults**

Ronan McGarrigle1, Sarah Knight1, Lyndon Rakusen1, Jason Geller2, & Sven Mattys1

*1Department of Psychology, University of York, UK.*

*2Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA, USA.*

This research was supported by an ESRC New Investigator Award (ES/R003572/1) to Ronan McGarrigle. The full study data set can be accessed at <https://osf.io/m42wq/>.

We sincerely thank all of the participants for kindly taking time out of their day to take part in the study. We would also like to thank everyone at SR research for their assistance with designing the experiments on E-Builder as well as everyone in the Speech lab at the University of York for useful discussions and insights.

Correspondence concerning this article should be addressed to:Ronan McGarrigle, Department of Psychology, University of York, UK, YO10 5DD. Email*:* [ronanomcg@gmail.com](mailto:ronanomcg@gmail.com).

**Abstract**

Listening to speech in adverse conditions can be challenging and effortful, especially for older adults. This study examined age-related differences in effortful listening by recording changes in the task-evoked pupil response (TEPR; a physiological marker of listening effort) both at the level of sentence processing and over the entire course of a listening task. A total of 65 (32 young adults; 33 older adults) participants performed a speech recognition task in the presence of a competing talker, while moment-to-moment changes in pupil size were continuously monitored. Participants were also administered the Vanderbilt Fatigue Scale; a questionnaire assessing daily life listening fatigue within four domains (social, cognitive, emotional, physical). Normalized TEPRs were overall larger and more steeply rising and falling around the peak in the older versus the young adult group during sentence processing. Additionally, mean TEPRs over the course of the listening task were more stable in the older versus the young adult group, consistent with a more sustained recruitment of compensatory attentional resources to maintain task performance. No age-related differences were found in terms of total daily life listening fatigue; however, older adults reported higher scores than young adults within the social domain. Overall, this study provides evidence for qualitatively distinct patterns of physiological arousal between young and older adults consistent with age-related upregulation in resource allocation during listening. A more detailed understanding of age-related changes in the subjective and physiological mechanisms that underlie effortful listening will ultimately help to address complex communication needs in aging listeners.

Keywords*:* Effortful listening, fatigue, pupillometry, aging, speech perception

**Introduction**

Speech perception is a fundamental building block of successful communication that relies heavily upon intact sensory and cognitive functions. Our cognitive systems take on increasing importance during speech perception in suboptimal conditions, e.g., in the presence of background noise (Rönnberg et al., 2008). Suboptimal listening conditions are a ubiquitous feature of everyday life and may present in a variety of forms and contexts (Mattys et al., 2012). As a result, the ability to understand speech is a non-trivial challenge for individuals with compromised hearing and/or cognitive ability. The natural process of aging is associated with declines in both auditory and cognitive functioning (Craik & Bialystok, 2006; Peelle & Wingfield, 2016). For example, older adults typically show increased outer hair cell loss in the cochlea and global reductions in grey matter in the auditory cortex (Peelle & Wingfield, 2016). Key cognitive skills, such as the ability to inhibit irrelevant information, also show decrements as a function of age (Borghini & Hazan, 2018; Braver & Barch, 2002; Hasher & Zacks, 1988; Tun et al., 2002). The ability to inhibit concomitant distractors (e.g., noise or a competing talker) while attending to speech is a particularly crucial skill for successful communication. Transient or more sustained difficulty with executing this skill may partly underlie reported difficulties with speech understanding associated with aging (Heinrich et al., 2016).

The functional consequences of an impoverished sensory and/or cognitive system are not limited to those which can be observed in a speech recognition accuracy test score. Indeed, the cognitive and physiological costs of achieving accurate speech perception are often missing from routine audiological and/or speech test examinations, but may impact one’s quality of life in profound ways (Phillips, 2016; Smith et al., 2011). Interest in quantifying the ‘effort’ associated with listening in adverse conditions led to the formation of the Framework for Understanding Effortful Listening (‘FUEL’); a consensus paper which defines listening effort as ‘the deliberate allocation of resourcesto overcome obstaclesin goal pursuit when carrying out a listening task’ (Pichora-Fuller et al., 2016). The concept of effort has traditionally been linked with fluctuations in autonomic arousal and particularly the ‘intensity’ with which one allocates attentional resources (Kahneman, 1973; Pichora-Fuller et al., 2016; Wingfield, 2016). It is thought that physiological indices may help not only to elucidate the underlying mechanisms of listening effort, but also to shed light on the relationship between listening effort and fatigue (Francis & Love, 2020). Fatigue from listening is a common complaint from individuals with hearing loss, although the precise relationship between transient manifestations of effort and longer-term fatigue remains elusive (Alhanbali et al., 2017; Hornsby et al., 2016; McGarrigle et al., 2014). A physiological measure that may help to elucidate the underlying mechanisms of effortful listening and fatigue is pupillometry (McGarrigle et al., 2017; Wang et al., 2018). Cognitive-evoked changes in pupil size have been shown to mirror activity in the brainstem’s locus coeruleus (Murphy et al., 2014), which is thought to play a central role in governing ongoing states of attention and arousal for goal-directed behavior (Aston-Jones & Cohen, 2005). In particular, the task-evoked pupil response (‘TEPR’) has been shown to be sensitive to changes in both the acoustic and cognitive demands of a listening task (Koelewijn et al., 2012; Kuchinsky et al., 2013; Winn et al., 2015; Zekveld et al., 2010).

Zekveld et al. (2011) examined TEPRs in a group of older adults with and without hearing loss. Data from a previous study on young adults with normal hearing were also analyzed for comparative purposes (Zekveld et al., 2010). They found that older adults with hearing loss showed a smaller reduction in TEPRs when listening in more favorable signal-to-noise ratios (SNRs) compared to both of the normal-hearing groups (one young adult and one age-matched). The smaller reduction in TEPR for the older adults with hearing loss was interpreted as reflecting less ‘release from effort’ when listening in conditions with increased intelligibility (Zekveld et al., 2011). Piquado et al. (2010) also investigated age differences in TEPR, this time during memory and sentence processing tasks. TEPRs were scaled according to each individual’s pupillary dynamic range to account for age-related differences in pupil reactivity (Bitsios et al., 1996; Tekin et al., 2018). Older adults showed larger normalized TEPRs than young adults when tasked with memorizing a series of digits for recall (regardless of digit list length) and also during the retention window of a sentence recognition task. The authors interpreted these findings as reflecting a global increase in resource allocation for older adults during language processing and the performance of a memory task (Piquado et al., 2010).

Age-related change may not be characterized solely by differences in overall levels of arousal and effort to support listening. Rather, qualitatively distinct patterns of attention and arousal during listening may reflect compensatory strategies in aging listeners (Wingfield & Grossman, 2006). For example, shifts in neural activation patterns from core primary auditory regions to areas associated with top-down cognitive control have been found in older adults when listening in degraded acoustic conditions (Eckert et al., 2016; Erb & Obleser, 2013; Peelle et al., 2010). In a study by Sharp et al. (2006), participants performed listening tasks that engaged both semantic and phonological language processing. Brain activity was monitored within the lateral prefrontal cortex (PFC) which is associated with attention and cognitive control (Shallice, 2002). A differential pattern of change over time in right lateral PFC activity was observed between young and older adult listeners over four blocks of trials. While young adults showed a reduction in PFC activity over time, older adults showed the reverse pattern. The authors suggested that this may reflect a compensatory mechanism used to maintain required levels of performance over time (Sharp et al., 2006). This pattern of data is consistent with the idea that older adults show more widely-distributed sensory and attention-related neural activation during task performance (Cabeza, 2002; Reuter-Lorenz & Cappell, 2008).

The deployment of top-down compensatory strategies to meet the demands of a degraded listening condition may manifest as a more sustained pattern of elevated physiological arousal and attention in older adults. Indeed, larger TEPRs during an effortful listening task have been shown to be associated with more vigilant behavioral response patterns during a continuous monitoring task in a group of older adults (Kuchinsky et al., 2016). Thomson and Hasher (2017) also examined age-related differences in sustained attention during the performance of two semantic vigilance tasks. Participants were asked to rate their subjective levels of arousal and effort following each task. They found no effect of age on task performance accuracy. However, older adults reported higher levels of subjective arousal (Expts 1 & 2) and effort (Expt 2) than the young adult group following task performance. The authors concluded that, while sustained vigilant attention was relatively preserved in older adults, this appeared to come at the cost of increased subjective arousal and effort (Thomson & Hasher, 2017).

While vigilance and sustained attention abilities are remarkably stable in older adults (Carriere et al., 2010), the physiological mechanisms that underlie such performance maintenance are not yet fully understood. It is possible that the need to sustain levels of attention and arousal during listening may at least partly underlie reports of tiredness from listening in aging listeners, thus providing a potential mechanistic link between transiently-measured effort during listening and more chronic incidences of fatigue. However, the relationship between effort and fatigue can also be thought of as bi-directional. The functional impact of fatigue may present in such a way that individuals who experience elevated fatigue will be unable to exert and sustain effort during a listening task, such that overall levels of effort during task performance will be relatively low (Wang et al., 2018). Indeed, Hess’ (2014) selective engagement theory of cognitive functioning in older adults posits that older adults are more likely to avoid effort expenditure during the performance of a challenging task to conserve their limited resources.

To our knowledge, no study to date has examined lab-based measures of effortful listening and daily life reports of fatigue from listening in healthy older adults. Similarly, few studies have sought to specifically investigate patterns of change over time in physiological arousal that support listening in adverse conditions for older adults. The present study aimed to uncover potentially differential patterns of attention and arousal in young versus older listeners during the performance of a sentence recognition task in the presence of a competing talker in a more positive SNR (‘easy’) condition and a more negative SNR (‘hard’) condition. In particular, we used Growth Curve Analysis (GCA) - a multi-level modelling approach - to more precisely characterize changes in the shape and timing of the TEPR at the individual trial level and across the duration of the listening task (Mirman, 2016). GCA has become widely popular as a way to capture potentially non-linear trajectories in the TEPR (Geller et al., 2019; McLaughlin & Van Engen, 2020). The current study also aimed to examine whether older adults differ from young adults in terms of their overall reports of daily life fatigue from listening as well as in their subjective experience of effort and fatigue during a listening task. Relative to a group of young adults, we predicted that older adults would show:

(1) Larger overall normalized TEPRs (i.e., a main effect of group) (Piquado et al., 2010)

(2) A smaller difference in TEPRs between ‘easy’ and ‘hard’ conditions (i.e., a Group x Condition interaction), consistent with less ‘release from effort’ in more favorable listening conditions (cf. Zekveld et al., 2011)

(3) A less steep decrease over time in TEPRs to reflect a more sustained pattern of arousal, supporting the age-related neural compensation hypothesis (Sharp et al., 2006)

(4) Higher self-report effort and tiredness from listening ratings and/or a more pronounced increase over time in tiredness from listening ratings, reflecting the heightened experience of fatigue due to listening demands

(5) Higher overall daily life listening fatigue scores

**Method**

Sample size, experimental design, hypotheses, outcome measures, and analysis plan for the experiment were pre-registered on the Open Science Framework (<https://osf.io/mjcn9>). Raw data and R scripts for analyses and plots can be found at <https://osf.io/m42wq/> along with stimuli used for the older adult participants. The stimuli used for young adult participants were the same as those used for a previous study, which can be found here <https://osf.io/cdv2r/>. Methods overlap considerably with a previous study conducted on young adults in the same lab (McGarrigle et al., 2020).

**Participants**

Participants were 32 young adults (2 male), aged 18-24 years (mean = 20, SD = 1) and 33 older adults (13 male), aged 62-82 years (mean = 68, SD = 5). We preregistered a target sample size of 28 participants in each group based on the rule of thumb of 1600 data points per condition for mixed model analyses (Brysbaert & Stevens, 2018)[[1]](#footnote-1). Data from an additional two older adult participants were excluded due to having cognitive assessment scores below the designated cut-off threshold (details provided below). All participants were native (British) English speakers who reported: (i) normal or corrected-to-normal visual acuity, (ii) no known eye condition, (iii) no history of suffering from claustrophobia (due to space restrictions in the testing booth) or any medical condition that could make them tired (e.g., Chronic Fatigue Syndrome, sleep disorder), and (iv) not currently a regular hearing aid user. Pure Tone Audiometry testing was conducted following the British Society of Audiology recommended procedure (2011) at 0.5, 1, 2, and 4 kHz in each ear[[2]](#footnote-2).

Older adult participants were screened for mild cognitive impairment using the Montreal Cognitive Assessment (MoCA) scale (Nasreddine et al., 2005). Based on the most recent guidelines for low false positive rate and better diagnostic accuracy (Carson et al., 2018), participants with a score of ≤ 23/30 were excluded from the analyses (*n* = 2). Young adult participants were recruited either through flyers posted around the University of York campus or as part of a course credit scheme for Psychology undergraduate students. Participants who did not receive course credit were financially compensated for their time. Older adults were recruited via the University of York participant pool, flyers posted in various public spaces around York (e.g., local library), or via an advertisement posted in the local newspaper. These participants were compensated for their time and fully reimbursed for any transport costs. All participants provided informed written consent before participating in the experiment. The study was granted ethical approval by the departmental research ethics committee at The University of York (ID: 733).

**Equipment**

A Kamplex Diagnostic Audiometer AD 25 was used for pure tone audiometry testing. For all subsequent testing, participants were positioned in a sound-treated test booth 65 cm away from a 24” flat screen LCD monitor, which displayed the visual stimuli. The participant’s head was stabilized on a head- and chin-rest which was secured to the end of a table. Stimulus presentation was programmed using the SR Research Experiment Builder software, version 2.2.1 (SR Research, Mississauga, ON, Canada). Auditory stimuli were presented via two speakers positioned either side of the computer monitor, at 45˚ and 315˚ azimuth angle. A microphone was positioned inside the test booth so that verbal responses could be heard and scored online by the experimenter who listened via headphones. Responses were also recorded for later verification of speech recognition scores.

Pupil size was recorded using the Eyelink 1000 Plus, at a sampling rate of 250 Hz. Pupil size was recorded as an integer number corresponding to the number of thresholded pixels in the camera’s pupil image. Typical pupil area can range between 100 and 10,000 units, with a precision of 1 unit. This corresponds to a resolution of .01 mm for a 5 mm pupil diameter. The desktop-mounted eye tracker camera was positioned in between the participant and the computer monitor at a distance of 55 cm from the participant (at 0˚ azimuth angle), following SR research guidelines for accurate and reliable tracking of the eye (SR Research, Mississauga, ON, Canada). The eye tracker camera was aligned to the center of the computer monitor screen, and was positioned just below the bottom of the flat screen to maximize the trackable range without obscuring the participant’s view of the screen.

**Materials**

Target speech stimuli were Harvard sentences (Rothauser et al., 1969) produced by a male talker with a standard Southern British accent. Each sentence contained five key words. The masker stimulus was a female talker, also with a Southern British accent, reading an extract from the standard phonetically-balanced ‘Rainbow Passage’ (Fairbanks, 1960). Target and masker stimuli were digitally mixed using a Matlab script (Nike, 2020) to create a series of .wav files at different SNRs. For each target sentence, a random 6-second portion of the masker audio file (total file duration: 74 seconds) was selected for target-masker mixing. Mixed target-and-masker .wav files were subsequently used for the adaptive screening and listening task (described below). For each trial, masker onset began two seconds before target onset and ended two seconds after target offset. Target stimulus presentation level was fixed at 55 dB SPL.

*Adaptive screening*

The adaptive screening used an approach similar to the one-up one-down adaptive procedure used to estimate 50% speech recognition performance accuracy (Kaernback, 1991). The purpose of this screening procedure was to calculate an SNR that could be used as the more challenging (hard) listening condition in the subsequent listening task (described in the next section). A performance criterion threshold of 50% correct was chosen as it has been shown to elicit the maximum TEPR (Ohlenforst et al., 2017). Twenty Harvard sentences were used for the adaptive screening.

For older adults, each Harvard sentence was mixed with the masker stimulus to create 15 different SNRs ranging from -10 dB to +4 dB SNR. Based on pilot testing, this SNR range was deemed suitable for capturing the potentially wide variability in the older adult speech recognition performance. This resulted in the creation of a total of 300 mixed target-masker .wav files (20 sentences x 15 SNRs). Older adult participants heard 20 mixed target and masker sentences, which started at +4 dB SNR and could reach a lower limit of -10 dB SNR. If participants responded correctly, the SNR decreased by 1 dB in the subsequent trial. If participants responded incorrectly, the SNR increased by 1dB in the subsequent trial. An incorrect response at the upper limit (i.e., +4 dB) or a correct response at the lower limit (i.e., -10 dB) resulted in no change to the SNR in the subsequent trial (i.e., it remained at +4 dB or -10 dB, respectively). Each participant’s 50% performance threshold was calculated as the mean SNR achieved in trials 15-20 of the adaptive procedure (rounded to the nearest whole number). For example, in cases where a ‘.5’ decimal value was calculated, we rounded down (e.g., -9.5 dB SNR was rounded down to -10 dB SNR). This adaptive approach was implemented to ensure that the hard condition was sufficiently challenging to require increased cognitive resource allocation, but not so challenging that it risked task disengagement (Borghini & Hazan, 2018).

For young adults, the adaptive screening process was the same as above, with the following exceptions: (1) 10 SNRs, ranging from -15 to -6 dB, were used to reflect superior speech recognition performance thresholds and reduced performance variability in young adult participants (McGarrigle, et al., 2020), resulting in a total of 200 target-masker .wav files (20 sentences x 10 SNRs), and (2) due to the reduced number of possible SNRs (i.e., 10 versus 15), 50% performance thresholds were this time calculated as the mean SNR in trials 10-20 (rather than 15-20). Overall, mean adapted SNR values were significantly different between groups (*t*(62) = 7.37, *p* < .001), with the young adult group achieving the 50% performance criterion at a more negative SNR (*M =* -7.9, *SD* = 1.20) than the older adult group (*M* = -3.10, *SD* = 3.50)[[3]](#footnote-3).

*Listening task*

The SNRs used during the listening task were individually adapted according to each participant’s performance during the adaptive screening. Mean SNR in the adaptive screening was used as the fixed hard condition SNR. The easy condition SNR was calculated as the hard condition SNR plus 10 dB. For example, a hard condition SNR of -6 dB would result in an easy condition SNR of +4 dB. A total of 120 Harvard sentences were used to create two target-masker lists (List 1 and List 2). Harvard sentences presented during the listening task differed from those presented in the adaptive screening. For List 1, 60 Harvard sentences were digitally mixed with the masker stimulus to create target-masker .wav files in all possible SNRs (15 for older adults, 10 for young adults) for the easy condition. Another 60 Harvard sentences were then digitally mixed with the masker stimulus to create a total of 60 target-masker .wav files in all possible SNRs for the hard condition. For List 2, the same 120 Harvard sentences were used, but the condition in which they appeared was swapped. Thus, if the target sentences in List 1 were used in the hard condition, the List 2 sentences were used in the easy condition, and so on. An additional four Harvard sentences were mixed with the masker stimulus to create practice trials at pre-determined easy and hard SNRs for each participant.

*Subjective ratings*

During the listening task, participants were administered self-report rating scales to assess: (1) tiredness from listening, (2) effort, and (3) performance evaluation. The three questions were as follows:

1. How tired of listening do you feel? (100-step scale from Not at all – Extremely)
2. How hard did you have to work to understand what was said for the previous 5 sentences? (100-step scale from Not at all – Extremely)
3. How would you rate your performance accuracy on the previous 5 sentences? (100-step scale from Poor – Good)

The choice of wording for the *tiredness from listening* scale was taken from Picou, Moore, and Ricketts (2017) and was chosen to tap tiredness arising specifically from listening demands, as opposed to other unrelated processes (e.g., visual fatigue). This measure has been shown to have high test-retest reliability (*r =* .84) and excellent internal consistency (*α =* .91) (Picou & Ricketts, 2018). The subjective *effort* rating scale was an adapted version of the NASA task load index item assessing mental demand (Hart & Staveland, 1988), to measure of effort (Dimitrijevic et al., 2019; McGarrigle et al., 2017; Strand et al., 2018). Finally, the subjective *performance evaluation* scale was an adapted version of the performance scales used in Moore and Picou (2018). This was included to help mitigate against the tendency for participants to use perceived performance as a proxy of effort (Moore & Picou, 2018). This measure was included in the analyses for exploratory purposes only.

Participants provided responses using an on-screen slider bar with values ranging from 0 to 100 in increments of 1. A triangular icon was positioned on the mid-point of the scale (50) to begin with and participants adjusted the icon using a mouse. Verbal anchors were positioned at each endpoint of the slider scale. A ‘Click here to continue’ box was positioned at the bottom of the screen which participants clicked on to advance to the next scale or trial.

*The Vanderbilt Fatigue Scale (VFS)*

The Vanderbilt Fatigue Scale (VFS) was administered to measure daily life experiences of listening-related fatigue. The VFS is designed to measure fatigue that is experienced specifically in the context of listening (see Bess et al., 2020 for a brief summary of the scale development process). While validated questionnaires for measuring general fatigue-related constructs do exist (e.g., Profile of Mood States; McNair et al., 1971), the VFS was deemed more suitable for the present study as it focuses specifically on fatigue that arises from communication challenges. The VFS consists of 40 items, each with 5-point Likert-type responses (0 – 4) with verbal anchors ranging from ‘Never/Almost never’ to ‘Always/Almost always’ or ‘Strongly Disagree’ to ‘Strongly Agree’. Examples of test items include; *‘I feel worn out from everyday listening’* and *‘It takes a lot of energy to listen and understand’*. One item (*‘I need to remove or turn off my hearing device to take a break from listening’*) is relevant only for individuals who wear a hearing device and as such was not included in the current study, leaving a total of 39 questionnaire items.

**Design and Procedure**

The same testing protocol was administered to both young and older adult participants, with one exception: on arrival, older adults were firstly taken to a quiet room for MoCA testing. MoCA testing involved participants completing a series of cognitive tests, including pencil and paper (e.g., copying the illustration of a cube) and verbal (e.g., digit rehearsal) tasks, all of which were supervised and led by the experimenter. The MoCA testing session lasted no longer than 10 minutes in total. MoCA scores were used for screening purposes only, and were therefore not included in the analyses. From this point onwards, the procedure was identical for both young and older adults.

Participants were seated comfortably in the testing booth and were then administered a pure tone audiometry test. After this test, eye tracker setup and calibration began. Soft room lighting was used and the computer screen had a grey background with reduced brightness settings (luminance from chin rest position measured at 100 cd/m2) to minimize any visual discomfort. The seat height and/or chinrest were adjusted to ensure that the participant was comfortable and their eyes were in line with the upper third of the screen. A 5-point calibration procedure was performed and subsequently validated. Participants were then given the following instructions prior to the adaptive screening task: *‘At the beginning of each trial, a black cross will be displayed on the screen. You will then hear an audio recording of a female talker and a male talker. The female talker will begin speaking before the male talker. Please continue to look at the black cross while you listen. After listening to the speech, text will be displayed on the screen asking you to respond. When prompted to do so, please repeat back the speech from the male talker only. If you are unsure what he said, please feel free to have a guess.’*

Participants performed 20 trials during the adaptive screening, starting at a SNR of +4 dB (or -6 dB for young adults). Participants began each trial by fixating on a small black cross in the center of the screen. The experimenter was seated outside the test booth and used a wireless keyboard to control stimulus presentation. For each trial, the experimenter pressed ‘y’ or ‘n’ on the keyboard to indicate whether the verbal response was correct or not (‘y’ = yes, ‘n’ = no). Participants could only advance to the next trial once the experimenter provided a keyboard response. A sentence was scored as correct only if all five key words were correctly identified and in the correct order. For example, for the sentence ‘The birch canoe slid on the smooth planks’, participants were only scored as correct if they accurately recalled all five key words in the correct order (i.e., birch, canoe, slid, smooth, planks). The adaptive screening lasted approximately 5 minutes.

After performing the adaptive screening, participants were then exposed to ‘bright’ and ‘dark’ settings to determine each participant’s pupillary dynamic range[[4]](#footnote-4). This would later be used to calculate a normalized TEPR that accounts for age-related differences in pupil size and responsivity (details provided in ‘Analysis’ section below). For the bright condition, the screen brightness was adjusted from 30% to 100% and all lights were switched on, including the lights outside and inside the testing booth. Luminance from the chin rest position in the ‘bright’ condition was measured at 335 cd/m2. Participants were asked to look at a fixation cross presented in the middle of the screen for 10 seconds, while their pupil size was recorded continuously. For the dark condition, all lights were switched off and the screen brightness was adjusted from 100% to 0%. Luminance from the chin rest position in the ‘dark’ condition was measured at 0.15 cd/m2. Once again, pupil size was recorded continuously for a period of 10 seconds. Because the screen was completely dark during this time, participants were simply asked to look toward the middle of the screen where the fixation cross had appeared in the previous ‘bright’ condition.

At the beginning of the listening task, participants were informed of the approximate task duration and that they would be asked to respond to subjective rating scales at periodic intervals. To familiarize themselves with the subjective rating scales, four practice trials (two in the easy SNR and two in the hard SNR) were performed. Listening task stimuli were presented in a blocked fashion. Easy and hard conditions contained 60 trials each. To avoid order effects, the order of the two SNR conditions was counterbalanced, with half of participants in each group performing the easy condition first (and the remaining half performing the hard condition first). Before each condition, participants provided a *tiredness from listening* scale response (used as a baseline in the analysis). *Effort* and *performance evaluation* rating scales were administered after every 5 trials (totaling 12 responses per participant per condition). The *tiredness from listening* scale was administered after every 10 trials (totaling 6 responses per participant per condition). At the relevant trial intervals, the *effort* scale was always administered first, followed by the *performance evaluation* scale and the *tiredness from listening* scale (in that order).

A three-second inter-trial interval (ITI) was incorporated in between the experimenter’s keyboard response to advance to the next trial and the onset of the female talker masker. Thus, including the experimenter's scoring time (~1 sec), there was at least 4-5 seconds between the participant’s verbal response and the recording of the subsequent trial baseline. This is consistent with Winn et al.’s (2018) recommended ITI of 4-6 seconds for experiments involving verbal responses. Each listening condition block lasted approximately 20 minutes. In between easy and hard conditions, participants were given the opportunity to briefly rest inside the booth or use the bathroom in the adjacent room.

After the listening task, participants were given the opportunity to have a 5-minute break. Once they were ready to continue, participants were provided with the VFS and a demographics questionnaire, which were completed on pen and paper. Written instructions for completing the VFS were provided at the beginning of the printed document, and the experimenter remained present to answer any questions. Figure 1 shows a schematic outline of the study procedure, which in total lasted around one hour and 10 minutes for young adults and one hour and 20 minutes for older adults.

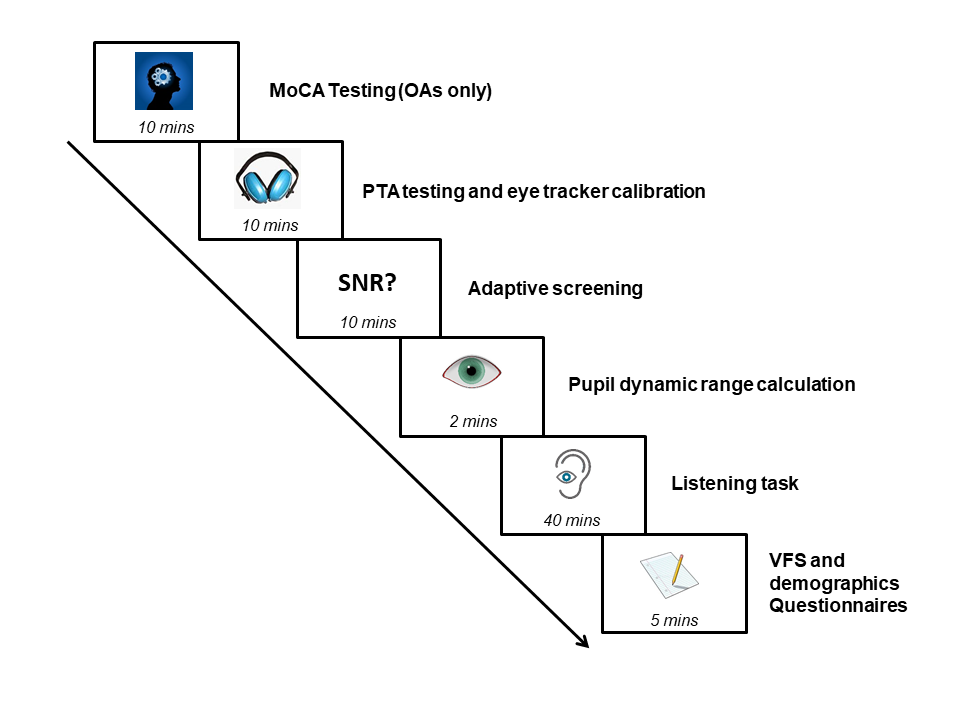


Figure 1. Schematic outline of the study procedure with time estimates for each component. MoCA = Montreal Cognitive Assessment; PTA = Pure Tone Audiometry; SNR = Signal-to-noise ratio; VFS = Vanderbilt Fatigue Scale.

**Analysis**

*Speech recognition performance*

Speech recognition performance was calculated as the mean percentage of key words correctly identified. Each trial contained five possible key words. The experimenter transcribed the responses online during the task. These scores were later verified offline by the experimenter and an independent rater using audio recordings of each trial. A three-way mixed ANOVA was conducted to examine the effects of group (young adult, YA; older adult, OA) and condition (easy, hard) on overall speech recognition performance and linear change in performance over time, with block (1, 2, 3, 4, 5, 6) modelled as a continuous variable. To rule out order effects, we included condition ‘order’ (easy-hard, hard-easy) as a covariate.

*Pupillometry*

**Pre-processing.**The R script used for pre-processing pupil data can be located at <https://osf.io/m42wq/>. Pupil data were pre-processed to remove noise from the analysis. Following data collection, a sample report was generated that included the pupil data for each participant and each trial. Gaze position is shown to influence pupil size estimation (Brisson et al., 2013). Therefore, to limit the influence of pupil size estimation errors caused by a rotated pupil (e.g., when looking at the corner of the screen), a rectangular area of interest was created in the center of the visual display surrounding the fixation cross (left, top, right, and bottom screen coordinates: 131, 94, 874, and 675, respectively). Only data from fixations that fell inside this perimeter were included in the sample report. These data were then output as a text file and read into R Studio (R Development Core Team, 2011) for pre-processing and analysis.

First, missing values in the data file (e.g., caused by blinks) were coded as ‘NA’. To account for spurious values caused by the eyelid closing and re-opening immediately before and after a blink, sample windows that included missing data were extended by 100 ms either side of the blink, using the GazeR package (Geller et al., 2020). Missing data were then linearly interpolated across using values either side of the extended window. Trials that contained > 50% missing data were removed from the analysis. Data from two older adult participants (s15 & s17) and one young adult participant (s24) were removed altogether due to having less than 20/60 trials with < 50% missing data within a single condition. Due to a technical error with the tracking software, pupil data for an additional two older adult participants (s8 & s9) were lost. Of the remaining participants, a total of 71 trials (1.9% of trials) were removed from the young adult data set and a total of 127 trials (3.6% of trials) were removed from the older adult data set. As a result, the remaining sample entered into the pupillometry analyses included 31 young adults and 29 older adults.

‘Senile miosis’ refers to the phenomenon whereby our pupils become smaller and less responsive (e.g., to changes in light conditions) as we age (Bitsios et al., 1996). To account for these age-related physiological differences, previous studies comparing young and older adults have calculated TEPRs that are scaled to each individual’s pupillary dynamic range (Ayasse et al., 2017). Therefore, for any between-subject comparisons, analyses were conducted on normalized TEPRs[[5]](#footnote-5). Adopting a similar approach to Ayasse et al. (2017), we calculated ‘normalized TEPR’ using the following equation;

(*d*M – *d*min) / (*d*max – *d*min) x 100

where *d*M is the participant’s pupil size at any single time point, *d*min is their mean pupil size during the ‘bright’ condition (i.e., minimum pupil size), and *d*max is their mean pupil size during the ‘dark’ condition (i.e., maximum pupil size).

After all pupil traces were normalized, baseline-correction was performed on each trial. The two seconds of masker-only presentation preceding the onset of the target speech was used as the baseline window. The mean pupil size value recorded during this two-second window was then subtracted from every sample recorded after target speech onset to produce a normalized TEPR value. The TEPR typically emerges approximately one second after target onset and peaks approximately one second after target offset (Winn et al., 2018). As a result, TEPR was calculated as the relative change from baseline during the 3-second window following target speech onset. This helped to rule out any pupil size changes elicited by behavioral and/or preparatory motor responses

**Growth Curve Analysis.**Growth curve analysis (GCA) is a multi-level regression technique for analyzing time series data. GCA can be used to capture changes in the shape and timing of the pupil response over time by fitting orthogonal polynomial time terms to the data. Orthogonal polynomials are transformations of natural polynomials, making each one (e.g., linear and quadratic) independent of one another, with each polynomial term capturing a distinct functional form. The ‘intercept’ refers to the overall mean pupil response (i.e., ‘area under the curve’), the ‘linear’ term refers to the slope of the pupil response (with more positive values indicating a steeper slope), and the ‘quadratic’ term captures the curvature of the inflection point in the time series, with more positive values indicating a flatter shape (Mirman, 2016). GCA was implemented in R Studio using R version 4.0.0 (R Development Core Team, 2019) using the‘lme4’ package (Bates et al., 2015). Plots were also implemented in R Studio, using the ‘ggplot2’ package (Wickham, 2016). Specifically, we used GCA to examine: (1) the mean trial-level TEPR (‘Mean TEPR’), and (2) the trial-by-trial change in the mean TEPR over the course of the experimental session (‘TEPR change over time’).

For both sets of analyses, fixed effects (i.e., group, condition, and polynomial time terms) and random effects (i.e., error terms) were modelled to predict the TEPR (see Tables 1 & 2 for final model formula). We also included condition ‘order’ (easy-hard, hard-easy) as a fixed effect in our GCA models. Condition, group, and order variables were sum coded (-1, 1) to obtain accurate main (rather than simple) fixed effect estimates. The random effects structure was selected as the maximal random effects structure justified by the design (Barr et al., 2013). Parameter estimates were reported using maximum likelihood estimation. An effect was deemed significant if its removal from the model resulted in a significant decrease in the log-likelihood ratio, which is distributed as χ2 with degrees of freedom equal to the number of parameters added. P values were calculated using the z distribution as an approximation of the t distribution (Mirman, 2016).

For the ‘Mean TEPR’ analysis, ‘time’ was modelled as a continuous variable representing time (in ms) from speech onset. The quadratic component of the TEPR was the highest order term of theoretical interest as this functional form can capture potential differences in the overall trajectory and curvature around the peak pupil response during speech processing. We therefore did not attempt to add higher-order terms (cubic, quartic, etc.) to the model. The final random effects structure included both subject and item level variance components. Specifically, participants were allowed to vary for each fixed effect time term overall (i.e., intercept, linear, and quadratic) as well as within each condition. Items were also allowed to vary for each fixed effect time term and for the effect of condition on the intercept and linear time terms. Item-level random effects for the effect of condition on the quadratic term were initially included in the maximal model, but removed due to model convergence issues[[6]](#footnote-6). The final model syntax can be found in Table 1.

For the ‘TEPR change over time’ analysis, ‘trial number’ (0 – 60) was modelled as a continuous variable. Based on visual inspection of the data, it was apparent that the pattern of change did not follow a nonlinear trajectory (see Figure 4). As a result, we did not attempt to add any higher-order terms (quadratic, cubic, etc.) to the model. For the random effects structure, participants were allowed to vary on the intercept and linear terms as well as for both Condition x Time term interactions (i.e., the effect of condition on the intercept and linear terms)[[7]](#footnote-7). Final model syntax can be found in Table 2. The R script for both of the GCA models including the steps described above can be found at <https://osf.io/m42wq/>.

*Subjective ratings*

Subjective ratings of effort, performance evaluation, and tiredness from listening ranged from 0 to 100. Tiredness from listening ratings were subtracted from a baseline score recorded at the beginning of each condition. A three-way mixed ANOVA was conducted to examine the effects of group (YA, OA) and condition (easy, hard) on overall subjective ratings and their linear change over time, with block (1, 2, 3, 4, 5, 6) modelled as a continuous variable. Condition ‘order’ was once again added as a covariate to rule out order effects. Mean by-block scores were calculated by averaging the two scores provided within each 10-trial block. For example, the first two ratings (after trials 5 and 10) were averaged to reflect overall effort/performance evaluation rating in block 1. Rating scores on trials 15 and 20 were averaged to reflect overall effort/performance evaluation rating in block 2, and so on.

*Daily life fatigue from listening*

Daily life fatigue from listening was calculated as the summed total score on all VFS items. Data from one young adult participant was omitted as they did not provide a response on two of the questionnaire items, leaving data for a total of 31 young adults and 33 older adults. Total scores were based on summed scores from 39 items, resulting in a total score that ranged between 0 – 156. As well as providing a total score, the VFS can also be used to generate domain-specific scores for each of the following four categories: (1) Social, (2) Cognitive, (3) Emotional, and (4) Physical. For example, *‘Feeling tired from listening causes strain on my relationships’* pertains to the ‘social’ domain, whereas *‘I get frustrated when I have to put a lot of energy into listening’* pertains to the ‘emotional’ domain. Of the 39 items in total, 10 pertained to social, cognitive, and emotional domains. The VFS item excluded from the questionnaire (*‘I need to remove or turn off my hearing device to take a break from listening’*) fell within the physical domain, meaning that the physical category scores represented a summed total score of 9 items (ranging from 0 – 36). Social, cognitive, and emotional scores represented a summed total of 10 items (ranging from 0 – 40). Five independent t-tests were conducted to examine mean differences in total fatigue and each of the four individual domain scores (social, cognitive, emotional, physical) as a function of group (YA, OA). Given the exploratory nature of these additional domain-specific analyses, we opted not to apply any correction for familywise error rate.

**Results**

**Speech recognition performance**

Figure 2 displays mean speech recognition performance for young versus older adults in each condition as a function of block. Overall, there was a main effect of condition (*F*(1, 63) = 51.81, *p* < .001, partial η2 = .46), with mean speech recognition performance higher in the easy (*M* = 94.35, *SE* = 0.33) than the hard (*M* = 74.28, *SE* = 1.03) condition. There was also a significant main effect of group (*F*(1, 63) = 6.60, *p* = .01, partial η2 = .10), with overall mean speech recognition performance higher in the older adults (*M* = 85.82, *SE* = 0.82) than the young adults (*M* = 82.81, *SE* = 0.84) group. There was a significant main effect of block on the linear term, (*F*(1, 63) = 8.91, *p* = .004, partial η2 = .13), with mean speech recognition performance showing a general improvement over time. All two- and three-way interactions were non-significant (*p*s > .05).



Figure 2. Mean speech recognition performance (% correct with ± SE lines) for easy and hard conditions as a function of block. YA = young adults; OA = older adults.

**Pupillometry**

*Mean TEPR*

Figure 3 illustrates mean TEPR as a function of condition (easy, hard) and group (YA, OA). Table 1 provides the full GCA model summary output for the ‘Mean TEPR’ analysis. There was a significant main effect of condition on the intercept term, (*χ*2 (1, N = 60) = 21.63, *p* < .001), and linear term, (*χ*2 (1, N = 60) = 46.03, *p* < .001). Overall, participants showed a larger and more steeply sloping TEPR in the hard versus the easy condition. There was a significant main effect of group on both the intercept term, (*χ*2 (1, N = 60) = 4.06, *p* = .04), and the quadratic term, (*χ*2 (1, N = 60) = 17.65, *p* < .001). TEPRs during speech recognition for older adults were overall larger and showed more of a quadratic shape function (i.e., more steeply rising and falling around the peak) compared to young adults (see Figure 3).



Figure 3. Normalized mean TEPR (with ± SE lines) overlaid with the final GCA quadratic model fit (solid lines). These data are plotted as a function of time (in milliseconds) from target speech onset. YA = young adults; OA = older adults.

Table 1. Growth curve analysis model code and full summary output for ‘Mean TEPR’ analysis. ‘Lin’ and ‘Quad’ are shorthand for linear and quadratic polynomial time terms, respectively.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| R Code: |  |  | |  | |  |  |
| TEPR ~ | (Lin+Quad)\* | Condition \* | | Group + Order + | | # Fixed effects | |
|  | ((Lin+Quad)\* | Condition | | | Subject) + | | # Random effects | |
|  | (Lin+Quad+ | Lin:Condition + | | Condition |Item) | | # Random effects | |
| Term | Estimate | SE | df | | *t* | | *p* |
| (Intercept) | 4.25 | 0.58 | 86.08 | | 7.31 | | < .001\*\*\* |
| Linear | 10.62 | 1.19 | 67.04 | | 8.94 | | < .001\*\*\* |
| Quadratic | -3.67 | 0.43 | 66.65 | | -8.59 | | < .001\*\*\* |
| Condition | -1.27 | 0.26 | 84.71 | | -4.89 | | < .001\*\*\* |
| Group | 0.94 | 0.51 | 57.07 | | 1.82 | | .07 |
| Order | 0.52 | 0.41 | 57.23 | | 1.27 | | .21 |
| Linear:Condition | -2.55 | 0.46 | 74.67 | | -5.49 | | < .001\*\*\* |
| Quadratic:Condition | -0.21 | 0.23 | 58.59 | | -0.91 | | .36 |
| Linear:Group | 0.05 | 1.15 | 58.04 | | 0.04 | | .97 |
| Quadratic:Group | -1.74 | 0.41 | 58.06 | | -4.23 | | < .001\*\*\* |
| Condition:Group | 0.09 | 0.23 | 57.72 | | 0.37 | | .71 |
| Linear:Condition:Group | -0.04 | 0.43 | 57.68 | | -0.08 | | .93 |
| Quadratic:Condition:Group | 0.09 | 0.23 | 58.55 | | 0.38 | | .71 |

\**p* < .05, \*\* *p* < .01, \*\*\**p* < .001.

Note. P-values shown are calculated based on the Wald test statistic rather than the Likelihood Ratio Test discussed in the text. While both test results are asymptotically equivalent, p-values derived from Wald tests are thought to be less reliable for small-to-moderate sample sizes (Agresti, 2007).

*TEPR change over time*

Figure 4 illustrates the mean TEPR change over time for each condition and group. Table 2 provides the full GCA model summary output for the ‘TEPR change over time’ analysis. There was a significant effect of group on the linear term, *χ*2 (1, N = 60) = 10.68, *p* = .001. Young adults showed a more pronounced negatively sloping mean TEPR over the course of the listening task compared with older adults who showed a more stable (i.e., less negatively sloping) TEPR pattern (see Figure 4).

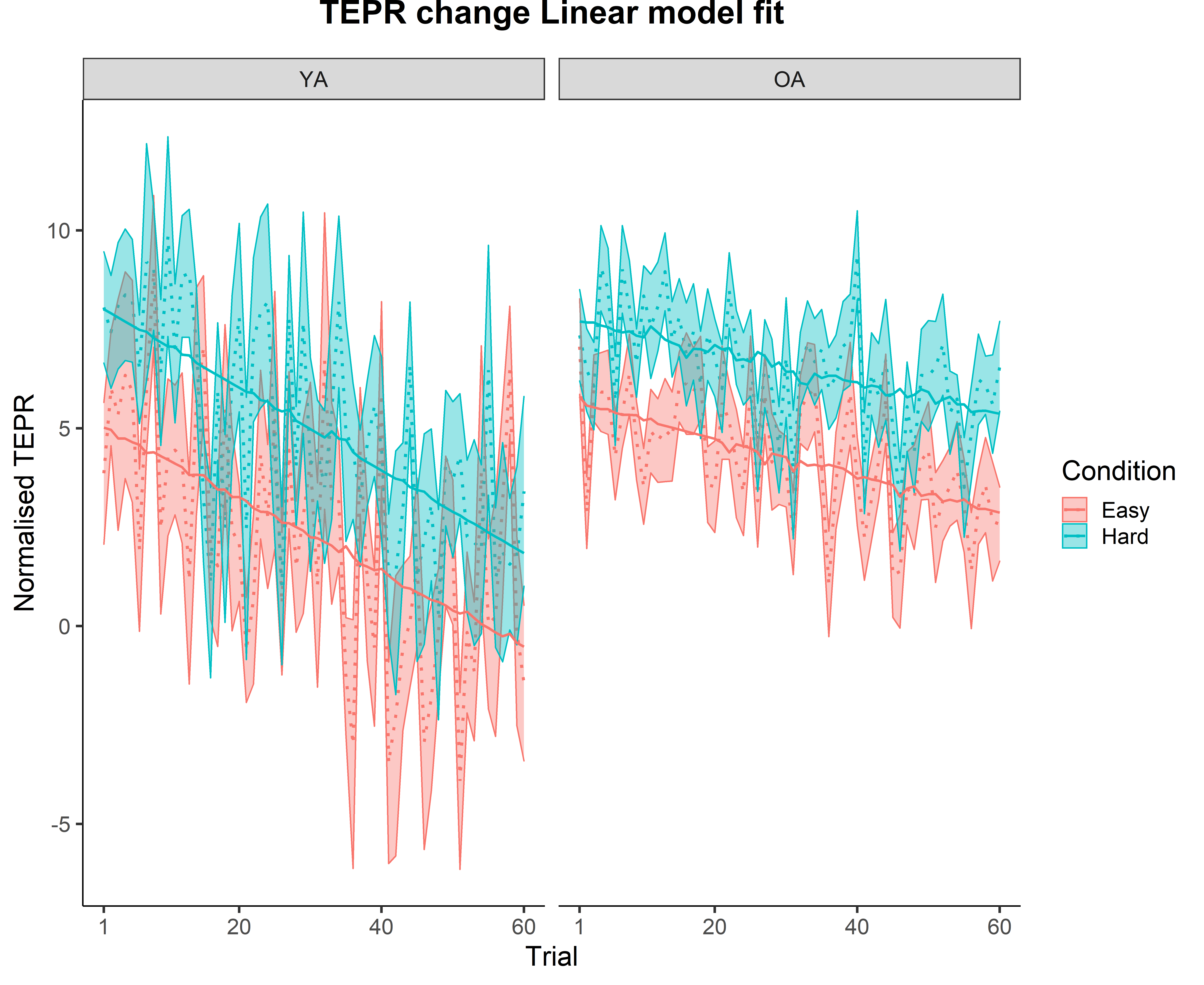
****

Figure 4. Normalized mean TEPR (with ± SE bands) overlaid with the final GCA linear model fit (solid lines). These data are plotted as a function of trial (1-60). YA = young adults; OA = older adults.

Table 2. Growth curve analysis model code and full summary output for ‘TEPR change over time’ analysis.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| R Code: |  |  |  |  |  |
| Mean trial TEPR ~ | (Linear)\* | Condition \* | Group + Order + | # Fixed effects |
|  | (Linear)\* | Condition | | Subject) | # Random effects |
| Term | Estimate | SE | df | *t* | *p* |
| (Intercept) | 4.72 | 0.71 | 58.76 | 6.70 | < .001\*\*\* |
| Linear | -9.50 | 1.32 | 58.32 | -7.19 | < .001\*\*\* |
| Condition | -1.26 | 0.24 | 58.72 | -5.29 | < .001\*\*\* |
| Group | -0.92 | 0.51 | 56.30 | -1.83 | .07 |
| Order | -0.49 | 0.98 | 57.26 | -0.50 | .62 |
| Linear:Condition | 0.15 | 0.85 | 164.63 | 0.18 | .86 |
| Linear:Group | -3.73 | 1.32 | 58.32 | -2.82 | .005\*\* |
| Condition:Group | -0.11 | 0.24 | 58.72 | -0.46 | .65 |
| Linear:Condition:Group | 0.63 | 0.85 | 164.64 | 0.73 | .46 |

Note. \**p* < .05, \*\* *p* < .01, \*\*\**p* < .001

**Subjective ratings**

Figure 5 displays each of the three subjective rating scores (effort, tiredness from listening, performance evaluation) for young versus older adults in each condition and block. There was a significant main effect of condition on mean effort ratings, (*F*(1, 63) = 43.81, *p* < .001, partial η2 = .41), with mean ratings higher in the hard (*M* = 66.61, *SE* = 1.89) than the easy (*M* = 32.08, *SE* = 2.26) condition. There was however no main effect of group, (*F*(1, 63) = .26, *p* = .61, partial η2 = .004),with mean effort ratings similar in young adults *(M* = 50.24*,* *SE* = 2.52) versus older adults (*M* = 48.44, *SE* = 2.48). There was no significant main effect of block on the linear term, (*F*(1, 63) = 2.73, *p* = .10, partial η2 = .04), with mean effort ratings showing no linear change over time. No significant two- or three-way interactions were found (all *p*s > .05).

We found a significant main effect of condition on tiredness from listening ratings, (*F*(1, 63) = 7.99, *p* = .006, partial η2 = .11), with mean ratings higher in the hard (*M* = 14.37, *SE* = 2.11) than the easy (*M* = 6.79, *SE* = 2.04) condition. However, there was no main effect of group, (*F*(1, 63) = 3.12, *p* = .08, partial η2 = .05),with mean tiredness from listening ratings not significantly different between young adults (*M* = 13.23, *SE* = 2.14) and older adults (*M* = 7.93, *SE* = 2.10). There was a significant main effect of block on the linear term, (*F*(1, 63) = 14.77, *p* < .001), partial η2 = .19, with tiredness from listening ratings showing a general increase over time. No significant two- or three-way interactions were found (all *p*s > .05).

Finally, there was a significant main effect of condition on performance evaluation ratings, (*F*(1, 63) = 46.53, *p* < .001, partial η2 = .43), with mean ratings higher in the easy (*M* = 69.87, *SE* = 1.94) than the hard (*M* = 35.01, *SE* = 1.90) condition. There was no main effect of group, (*F*(1, 63) = .93, *p* = .34, partial η2 = .02),with mean performance evaluation ratings similar in young adults (*M* = 50.95, *SE* = 2.21) and older adults (*M* = 53.93, *SE* = 2.17). There was also no main effect of block on the linear term, (*F*(1, 63) = .25, *p* = .62, partial η2 = .004), with performance ratings showing little change over time. No significant two- or three-way interactions were found (all *p*s > .05).

****

Figure 5. Mean subjective ratings (0-100 scale, with ± SE lines) for each condition (circles = easy; triangles = hard) as a function of block. Tiredness from listening ratings were calculated as the relative change from a baseline recorded at the beginning of block 1. YA = young adults; OA = older adults.

**Daily life fatigue from listening**

Total fatigue scores derived from the VFS showed high reliability (Cronbach’s α = 0.96). All four VFS subscales also showed high reliabilities (emotional and social subscales, Cronbach’s α = 0.87; cognitive and physical subscales, Cronbach’s α = 0.85). Total fatigue scores did not differ significantly between young adults (mean = 32.39, SD = 17.34) and older adults (mean = 37.55, SD = 22.47), *t*(62) = 1.02, *p* = .31, *d* = .26. Similarly, no significant differences were found between young and older adults within the emotional (YA, *M* = 8.19, *SD* = 5.44; OA, *M* = 8.33, *SD* = 5.72), physical (YA, *M* = 5, *SD* = 3.98; OA, *M* = 6.24, *SD* = 5.23), or cognitive (YA, *M* = 12.97, *SD* = 4.98; OA, *M* = 13.55, *SD* = 6.58) domains (all *ps* > 0.05). However, older adults showed significantly higher scores than young adults in the social domain (YA, *M* = 6.23, *SD* = 4.81; OA, *M* = 9.42, *SD* = 6.83), *t*(58[[8]](#footnote-8)) = 2.18, *p* = .03, *d* = .54. See Figure 6 for boxplots displaying total fatigue VFS scores and fatigue scores within the social domain.

****

Figure 6. Boxplots showing total VFS scores (left hand side) and social domain VFS scores (right hand side) as a function of group. YA = young adults; OA = older adults. Total fatigue scores range from 0 – 156. Social domain fatigue scores range from 0 – 40. Each data point reflects individual participant test scores, jittered for illustrative purposes.

**Discussion**

The current study examined the effects of age and SNR on physiological and subjective markers of effortful listening. We also investigated age-related differences in daily life fatigue from listening. Overall, we found an effect of condition on TEPRs in both young and older adult listeners; both groups showed larger and more steeply rising TEPRs in the hard versus the easy SNR condition. We also found age-related differences in the overall amplitude and shape of the TEPR during sentence processing; older adults showed a larger and more steeply rising and falling pupil dilation response compared to the relatively smaller and flatter TEPR in young adults. When examining change over time, mean TEPRs showed a more pronounced negative slope in the young versus the older adult group. This is consistent with the notion that performance was maintained over the course of the listening task at the expense of a relatively more sustained pattern of attention and arousal in older adult listeners. Older adults did not differ from young adults in terms of overall daily life fatigue from listening scores. However, exploratory analyses revealed significant difference between groups in a specific dimension of fatigue, with older adults reporting heightened levels of fatigue within the social domain.

**Effects of age and task demand on TEPR**

Overall, we found support for hypothesis (1); older adults showed larger normalized TEPRs than young adults during an effortful listening task. As illustrated in Figure 3, the older adult group TEPR was characterized by a larger and more steeply rising and falling peak response compared with the smaller and flatter TEPR shown in the young adult group. The contrasting TEPR amplitude and shape suggests that, on average, older listeners recruited additional resources during speech processing culminating in a larger peak response at verbal stimulus offset. This finding supports the idea that older listeners likely recruit compensatory resources to offset the cost of reduced sensory and/or cognitive acuity during speech recognition (Tun et al., 2010; Wingfield et al., 2005). The effect of group on the quadratic term captures the fact that the increased pupil response was not sustained and instead returned to a pupil size value comparable with the young adult group just before the upcoming sentence recall window. This illustrates the value of capturing not just the mean and peak response, but also the overall shape and timing of the pupil response (Geller et al., 2019; Kuchinsky et al., 2013; Mirman, 2016).

We did not find support for hypothesis (2); there was no Group x Condition interaction on normalized TEPRs. Both age groups showed a similar effect of SNR on mean normalized TEPR, with larger overall pupil dilation shown in the hard versus the easy condition. This suggests that a 10 dB decrease in SNR modulates listening effort in a manner that is comparable between both groups. However, it should be noted that the current study used a normalization procedure that was not used in Zekveld et al. (2011). Indeed, when we consider the pattern of results for non-normalized TEPRs (see supplementary file 2), the data are more consistent with Zekveld et al. (2011); in other words, the difference in TEPR as a function of acoustic task demand appears to be reduced in the older, versus the younger, adult group[[9]](#footnote-9). Further, the sample recruited in Zekveld et al. (2011) were younger on average than the older adult group in the current study (mean age = 61 years) and had presented with a clinical hearing loss (see Figure 1, Zekveld et al., 2011). Relatively poor hearing acuity may have contributed to the reduced impact of SNR on TEPR.

Finally, we found support for hypothesis (3). The young adult group showed a more steeply decreasing pattern of change in mean TEPRs than the older adult group over the course of the experimental task across both SNR conditions (see Figure 4). This finding is consistent with age-related neural compensation during a sustained cognitively-effortful task (Cabeza, 2002; Reuter-Lorenz & Cappell, 2008; Sharp et al., 2006; Wingfield & Grossman, 2006). We also extended Thomson and Hasher’s (2017) finding with complementary physiological evidence suggesting that relatively stable vigilance and attentional monitoring in older adults comes at the cost of increased autonomic arousal. Although we attempted to equate performance between groups as much as possible by using an adaptive procedure, speech recognition performance was nonetheless overall significantly higher in the older versus the young adult group during the experimental listening task (see Figure 2). In light of the relatively stable performance scores shown in the older adult group, the TEPR data are more consistent with the idea of age-related frontal compensation (i.e., use of attentional strategies to help optimize performance) than with age-related reductions in neural inhibition which may present in a similar fashion (Martins et al., 2015). A reduction in inhibitory brain activity would likely have manifested behaviorally as a decrement in speech recognition performance following distraction and/or lapses in concentration. However, additional correlation analyses revealed no significant association between mean TEPR and mean speech recognition performance in both groups (*p*s > .05). As a result, support for the age-related neural compensation hypothesis is tentative only.

**Effects of age and task demand on subjective ratings of effortful listening**

Both young and older adults showed a similar pattern of increased effort and tiredness from listening ratings in the hard versus the easy condition, and overlapping changes over time in tiredness from listening. We therefore did not find support for hypothesis (4) and cannot corroborate Thomson and Hasher’s (2017) findings of increased subjective effort and arousal ratings in older adults versus young adults. Although there was a numerical difference in tiredness from listening ratings between groups (see Figure 5), the main effect of group did not reach significance (*p* = .08). Therefore, although we found physiological evidence of a difference between groups in the effortful control of attention during listening, this did not translate into differences in perceived effort or tiredness from listening. Effort ratings have been shown to be highly correlated with subjective performance evaluation (Moore & Picou, 2018; McGarrigle, et al., 2020). It is therefore possible that attempting to equate performance between groups served to minimize potential differences in perceived effort and tiredness from listening. Another possibility is that any potential age-related differences in subjective ratings were curtailed by the tendency for older adults to underestimate the level of difficulty experienced during a listening task (Larsby et al., 2005).

**Daily life fatigue from listening**

We found only partial support for hypothesis (5); older adults did not differ from young adults in terms of overall daily life listening-related fatigue scores. However, exploratory analyses revealed a significant difference between groups in a specific dimension of fatigue, with the older adult group reporting increased levels of fatigue within the social domain. The VFS was designed to measure daily life fatigue from listening across a wide range of ages (Bess et al., 2020)[[10]](#footnote-10). Based on the negative impact of sensory and/or cognitive decline, we predicted that older adults would report higher overall levels of daily life fatigue from listening (Alhanbali et al., 2017, 2018). Contrary to our expectations, we found no significant difference in total fatigue scores between groups. There are a number of potential reasons for this. It is possible that the experience of fatigue from listening in the older adult population shows a high degree of between-subject variability. Indeed, the increased variability shown in older adults versus young adults lends some support to this assertion (see Figure 6; left plot). Fatigue from listening might therefore only be considered problematic for certain individuals, irrespective of age and/or hearing acuity. Recent studies investigating fatigue in hearing-impaired populations attest to the inherent variability in fatigue reports (Holman et al., 2019; Hornsby & Kipp, 2016). A second possible interpretation is that the tendency for older adults to underestimate the level of difficulty that they encounter during everyday listening may extend to the subjective experience of fatigue. Larsby et al. (2005) found that, despite markedly different behavioral task scores between young and older adults, older adults reported relatively less listening effort. This may be particularly true of communication-related difficulties which may be perceived as relatively less serious or life-threatening, meaning that older adults feel less inclined to ‘complain’. Further research is warranted to explore this possibility.

The VFS also provides scores pertaining to specific dimensions of fatigue, specifically; cognitive, social, emotional, and physical. We conducted a set of exploratory analyses to investigate whether there were any age-related differences within each of these domains. Young and older adults showed similar levels of fatigue within emotional, physical, and cognitive domains. However, the older adult group showed significantly higher levels of fatigue within the social domain (see Figure 6; right plot). It is possible therefore that the aspects of fatigue that are most salient to older adults are those that impact upon their social functioning (e.g., reluctance to attend a social event in suboptimal acoustic conditions). In contrast, this may be of relatively reduced importance for young adults who more regularly engage socially in listening conditions that would be considered suboptimal (e.g., cafes, bars, music gigs). However, given the exploratory nature of these domain-specific analyses, we would encourage due caution in their interpretation. Further research is warranted to verify this association and examine the nature of these potential age-related differences in more detail.

**Limitations and Future Directions**

One potential limitation of this study is that we are unable to shed light on the potential influence of time-of-day on performance and arousal patterns. Time of day has been found to influence circadian arousal rhythm and susceptibility to distraction in older adults (Anderson et al., 2014). All participants in the current study were tested in the hours between 11am – 4pm. It is therefore unlikely that there were systematic differences in circadian arousal rhythm between age groups. However, a more targeted examination of the effects of time-of-day is warranted in future research. Also, we interpret differences in arousal patterns between age groups to reflect more sustained effortful listening in older adults. However, we cannot rule out the possibility that this effect may be partly driven by an increase in more general task-related engagement during language processing (Davis et al., 2014). Indeed, the sentence transcription task administered in the current study relies on intact working memory for subsequent retrieval of the sentence which may place differentially greater demands on aging listeners.

**Conclusions**

The current study examined age-related differences in effortful listening and fatigue using a host of complementary subjective measures and a physiological marker of effort (TEPR). We revealed age-related differences in TEPR both at the sentence perception level and also over the course of a period of sustained effortful listening. Patterns of physiological arousal shown are consistent with the notion that listening in the presence of a competing talker comes at the cost of increased (and more sustained) mental effort as we age. A better understanding of the functional consequences of aging on everyday listening will ultimately help researchers and clinicians to tailor interventions that seek to optimize communicative success and well-being in older listeners.

**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

Anderson, J. A. E., Campbell, K. L., Amer, T., Grady, C. L., & Hasher, L. (2014). Timing is everything: age differences in the cognitive control network are modulated by time of day. *Psychology and Aging*, *29*(3), 648–657. https://doi.org/10.1037/a0037243

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annual Review of Neuroscience*, *28*, 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Ayasse, N. D., Lash, A., & Wingfield, A. (2017). Effort not speed characterizes comprehension of spoken sentences by older adults with mild hearing impairment. *Frontiers in Aging Neuroscience*, *8,* 329. https://doi.org/10.3389/fnagi.2016.00329

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). Parsimonious mixed models. *ArXiv:1506.04967 [Stat]*. http://arxiv.org/abs/1506.04967

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bess, F. H., Davis, H., Camarata, S., & Hornsby, B. W. Y. (2020). Listening-related fatigue in children with unilateral hearing Loss. *Language, Speech, and Hearing Services in Schools*, *51*(1), 84–97. https://doi.org/10.1044/2019\_LSHSS-OCHL-19-0017

Bitsios, P., Prettyman, R., & Szabadi, E. (1996). Changes in autonomic function with age: a study of pupillary kinetics in healthy young and old people. *Age and Ageing*, *25*(6), 432–438. https://doi.org/10.1093/ageing/25.6.432

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: a pupillometry study. *Frontiers in Neuroscience*, *12*, 152. https://doi.org/10.3389/fnins.2018.00152

Braver, T. S., & Barch, D. M. (2002). A theory of cognitive control, aging cognition, and neuromodulation. *Neuroscience & Biobehavioral Reviews*, *26*(7), 809–817. https://doi.org/10.1016/S0149-7634(02)00067-2

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods*, *45*(4), 1322–1331. https://doi.org/10.3758/s13428-013-0327-0

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: a tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Cabeza, R. (2002). Hemispheric asymmetry reduction in older adults: The HAROLD model. *Psychology and Aging*, *17*(1), 85–100. https://doi.org/10.1037//0882-7974.17.1.85

Carson, N., Leach, L., & Murphy, K. J. (2018). A re-examination of Montreal Cognitive Assessment (MoCA) cutoff scores: Re-examination of MoCA cutoff scores. *International Journal of Geriatric Psychiatry*, *33*(2), 379–388. https://doi.org/10.1002/gps.4756

Craik, F. I. M., & Bialystok, E. (2006). Cognition through the lifespan: Mechanisms of change. *Trends in Cognitive Sciences*, *10*(3), 131–138. https://doi.org/10.1016/j.tics.2006.01.007

Davis, S. W., Zhuang, J., Wright, P., & Tyler, L. K. (2014). Age-related sensitivity to task-related modulation of language-processing networks. *Neuropsychologia*, *63*, 107–115. https://doi.org/10.1016/j.neuropsychologia.2014.08.017

Dimitrijevic, A., Smith, M. L., Kadis, D. S., & Moore, D. R. (2019). Neural indices of listening effort in noisy environments. *Scientific Reports*, *9*, 11278. https://doi.org/10.1038/s41598-019-47643-1

Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdisciplinary Reviews. Cognitive Science*, *11*(1), e1514. https://doi.org/10.1002/wcs.1514

Geller, J., Landrigan, J.-F., & Mirman, D. (2019). A pupillometric examination of cognitive control in taxonomic and thematic semantic memory. *Journal of Cognition*, *2*(1), 6. https://doi.org/10.5334/joc.56

Geller, J., Winn, M. B., Mahr, T., & Mirman, D. (2020). GazeR: A package for processing gaze position and pupil size data. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01374-8

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Advances in Psychology* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/10.1016/S0166-4115(08)62386-9

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *Psychology of Learning and Motivation* (Vol. 22, pp. 193–225). Academic Press. https://doi.org/10.1016/S0079-7421(08)60041-9

Heinrich, A., Gagne, J.-P., Viljanen, A., Levy, D. A., Ben-David, B. M., & Schneider, B. A. (2016). Effective communication as a fundamental aspect of active aging and well-being: Paying attention to the challenges older adults face in noisy environments. *Social Inquiry into Well-Being*, *2*(1), 51–69. https://doi.org/10.13165/SIIW-16-2-1-05

Holman, J. A., Drummond, A., Hughes, S. E., & Naylor, G. (2019). Hearing impairment and daily-life fatigue: A qualitative study. *International Journal of Audiology*, *58*(7), 408–416. https://doi.org/10.1080/14992027.2019.1597284

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. https://doi.org/10.1097/AUD.0000000000000203

Hornsby, B. W. Y., Naylor, G., & Bess, F. H. (2016). A taxonomy of fatigue concepts and their relation to hearing loss. *Ear and Hearing*, *37*(Suppl 1), 136S-144S. https://doi.org/10.1097/AUD.0000000000000289

Kahneman, D. (1973). *Attention and Effort*. https://doi.org/10.2307/1421603

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker: *Ear and Hearing*, *33*(2), 291–300. https://doi.org/10.1097/AUD.0b013e3182310019

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*(1), 23–34. https://doi.org/10.1111/j.1469-8986.2012.01477.x

Kuchinsky, S. E., Vaden, K. I., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2016). Task-related vigilance during word recognition in noise for older adults with hearing loss. *Experimental Aging Research*, *42*(1), 50–66. https://doi.org/10.1080/0361073X.2016.1108712

Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology*, *44*(3), 131–143. https://doi.org/10.1080/14992020500057244

Martins, R., Joanette, Y., & Monchi, O. (2015). The implications of age-related neurofunctional compensatory mechanisms in executive function and language processing including the new Temporal Hypothesis for Compensation. *Frontiers in Human Neuroscience*, *9*, 221. https://doi.org/10.3389/fnhum.2015.00221

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*, *54*(2), 193–203. https://doi.org/10.1111/psyp.12772

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), 433–445. https://doi.org/10.3109/14992027.2014.890296

McGarrigle, R., Rakusen, L., & Mattys, S. (2020). *Effortful listening under the microscope: Examining relations between pupillometric and subjective markers of effort and tiredness from listening.* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/kych5

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America*, *147*(2), EL151–EL156. https://doi.org/10.1121/10.0000718

Mirman, D. (2016). *Growth Curve Analysis and Visualization Using R*. CRC Press.

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. https://doi.org/10.1044/2018\_JSLHR-H-17-0451

Murphy, P. R., O’Connell, R. G., O’Sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping*, *35*(8), 4140–4154. https://doi.org/10.1002/hbm.22466

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., & 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

Peelle, J. E., & Wingfield, A. (2016). The neural consequences of age-related hearing loss. *Trends in Neurosciences*, *39*(7), 486–497. https://doi.org/10.1016/j.tins.2016.05.001

Phillips, N. A. (2016). The implications of cognitive aging for listening and the framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*, 44S. https://doi.org/10.1097/AUD.0000000000000309

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). *Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL)*. *37*, 23.

Picou, E. M., Moore, T. M., & Ricketts, T. A. (2017). The effects of directional processing on objective and subjective listening effort. *Journal of Speech, Language, and Hearing Research*, *60*(1), 199–211. https://doi.org/10.1044/2016\_JSLHR-H-15-0416

Picou, E. M., & Ricketts, T. A. (2018). The relationship between speech recognition, behavioural listening effort, and subjective ratings. *International Journal of Audiology*, *57*(6), 457–467. https://doi.org/10.1080/14992027.2018.1431696

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, *47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Reuter-Lorenz, P. A., & Cappell, K. A. (2008). Neurocognitive aging and the compensation hypothesis. *Current Directions in Psychological Science*, *17*(3), 177–182. https://doi.org/10.1111/j.1467-8721.2008.00570.x

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

Shallice, T. (n.d.). Fractionation of the Supervisory System. In *Principles of Frontal Lobe Function*. Oxford University Press. Retrieved 22 April 2020, from https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780195134971.001.0001/acprof-9780195134971-chapter-17

Sharp, D. J., Scott, S. K., Mehta, M. A., & Wise, R. J. S. (2006). The neural correlates of declining performance with age: Evidence for age-related changes in cognitive control. *Cerebral Cortex (New York, N.Y.: 1991)*, *16*(12), 1739–1749. https://doi.org/10.1093/cercor/bhj109

Smith, S. L., Pichora-Fuller, K. M., Watts, K. L., & La More, C. (2011). Development of the listening self-efficacy questionnaire (LSEQ). *International Journal of Audiology*, *50*(6), 417–425. https://doi.org/10.3109/14992027.2011.553205

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

Tekin, K., Sekeroglu, M. A., Kiziltoprak, H., Doguizi, S., Inanc, M., & Yilmazbas, P. (2018). Static and dynamic pupillometry data of healthy individuals: Normative data for pupillometry. *Clinical and Experimental Optometry*, *101*(5), 659–665. https://doi.org/10.1111/cxo.12659

Thomson, D. R., & Hasher, L. (2017). On the preservation of vigilant attention to semantic information in healthy aging. *Experimental Brain Research*, *235*(7), 2287–2300. https://doi.org/10.1007/s00221-017-4969-5

Tun, P. A., Benichov, J., & Wingfield, A. (2010). Response latencies in auditory sentence comprehension: Effects of linguistic versus perceptual challenge. *Psychology and Aging*, *25*(3), 730–735. https://doi.org/10.1037/a0019300

Tun, P. A., O’Kane, G., & Wingfield, A. (2002). Distraction by competing speech in young and older adult listeners. *Psychology and Aging*, *17*(3), 453–467. https://doi.org/10.1037/0882-7974.17.3.453

Wang, Y., Naylor, G., Kramer, S. E., Zekveld, A. A., Wendt, D., Ohlenforst, B., & Lunner, T. (2018). Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. *Ear and Hearing*, *39*(3), 573. https://doi.org/10.1097/AUD.0000000000000512

Wingfield, A. (2016). Evolution of models of working memory and cognitive resources: *Ear and Hearing*, *37*, 35S-43S. https://doi.org/10.1097/AUD.0000000000000310

Wingfield, A., & Grossman, M. (2006). Language and the aging brain: Patterns of neural compensation revealed by functional brain imaging. *Journal of Neurophysiology*, *96*(6), 2830–2839. https://doi.org/10.1152/jn.00628.2006

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing loss in older adulthood: What it is and how it interacts with cognitive performance. *Current Directions in Psychological Science*, *14*(3), 144–148. https://doi.org/10.1111/j.0963-7214.2005.00356.x

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation: *Ear and Hearing*, *36*(4), e153–e165. https://doi.org/10.1097/AUD.0000000000000145

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, *22*, 233121651880086. https://doi.org/10.1177/2331216518800869

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

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb

1. This discrepancy reflects our anticipated attrition rate of approximately 10% due to the pupillometry data exclusion criteria (see details in the ‘Analysis’ section). More details about the power analysis can be found on OSF (<https://osf.io/mjcn9>). [↑](#footnote-ref-1)
2. See supplementary file (3) for mean pure tone audiometric thresholds for each group. [↑](#footnote-ref-2)
3. SNR values were stored in the same results file as the pupillometry data. As a result, these values were successfully recorded for all but one OA participant (s8), whose pupillometry data was completely lost due to a technical error with the eye tracker (reported also in the ‘Analysis’ section below). [↑](#footnote-ref-3)
4. Statistical analyses of the pupil dynamic range values in each group (OA, YA) and condition (bright, dark) are provided in the supplementary materials (supplementary file 1). [↑](#footnote-ref-4)
5. Results on non-normalized mean TEPR data are also provided in supplementary materials (supplementary file 2). [↑](#footnote-ref-5)
6. Model non-convergence suggests that the model may be over-parameterized and that the random effects structure is too complex (Bates et al., 2018). [↑](#footnote-ref-6)
7. Items were not included as random effects as there were an insufficient number of item list levels (i.e., there were only two possible item list: condition variations). [↑](#footnote-ref-7)
8. Levene’s test indicated unequal variances (F = 4.63, p = .04), so degrees of freedom were adjusted from 62 to 58. [↑](#footnote-ref-8)
9. Analysis conducted on the non-normalized data revealed a significant Group x Condition interaction on the intercept term. [↑](#footnote-ref-9)
10. Although nearing completion, please note that the VFS is still in the process of extensive scale development and validation. [↑](#footnote-ref-10)