Pupillometry reveals changes in physiological arousal during a sustained listening task

DOI:
10.1111/psyp.12772

Document Version
Accepted author manuscript

Citation for published version (APA):
McGarrigle, R., Dawes, P., Stewart, A., Kuchinsky, S. E., & Munro, K. (2016). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology*, *54*(2), 193-203. https://doi.org/10.1111/psyp.12772

Published in:
Psychophysiology

Pupillometry reveals changes in physiological arousal during a sustained listening task

Ronan McGarrigle,[a] Piers Dawes,[a] Andrew J Stewart,[b] Stefanie E Kuchinsky,[c,d] and Kevin J Munro [a,e]

[a]*Manchester Centre for Audiology and Deafness, School of Health Sciences, University of Manchester, UK.*

[b]*Division of Neuroscience and Experimental Psychology, School of Biological Sciences, University of Manchester, UK.*

[c]*Center for Advanced Study of Language, University of Maryland, USA.*

[d]*Maryland Neuroimaging Center, University of Maryland, USA.*

[e]*Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK.*

Corresponding author: Ronan McGarrigle, Department of Hearing and Speech Sciences, Vanderbilt University, Nashville, TN 37212, USA

Email: ronanomcg@hotmail.com

Physiological changes during sustained listening

Abstract

Hearing loss is associated with anecdotal reports of fatigue during periods of sustained listening. However, few studies have attempted to measure changes in arousal, as a potential marker of fatigue, over the course of a sustained listening task. The present study aimed to examine subjective, behavioural, and physiological indices of listening-related fatigue. Twenty-four normal-hearing young adults performed a speech-picture verification task in different signal-to-noise ratios (SNRs) while their pupil size was monitored and response times (RTs) recorded. Growth curve analysis (GCA) revealed a significantly steeper linear decrease in pupil size in the more challenging SNR, but only in the second half of the trial block. Changes in pupil dynamics over the course of the more challenging listening condition block suggest a reduction in physiological arousal. Behavioural and self-report measures did not reveal any differences between listening conditions. This is the first study to show reduced physiological arousal during a sustained listening task, with changes over time consistent with the onset of fatigue.

Descriptors: Listening-related fatigue, Mental fatigue, Pupillometry, Narrative speech processing, Growth curve analysis.

Fatigue is a multi-faceted construct that has many different causes (e.g., mental stress, overstimulation, boredom) and domains (e.g., mental and physical). In healthy individuals, fatigue is predictable and transient and typically occurs as a physiological reaction to prolonged and/or intense activity (Finsterer & Mahjoub, 2014). However, in certain chronic illnesses (e.g., Crohn's disease, chronic fatigue syndrome, multiple sclerosis), fatigue can also cause emotional distress and negatively impact quality of life (Bower, 2012; Finsterer & Mahjoub, 2014). Few studies have investigated fatigue as a consequence of hearing loss. A common complaint reported by individuals with hearing loss is the mental fatigue resulting from sustained effortful listening in challenging environments (Bess & Hornsby, 2014). Listening-related fatigue can generally be considered as a feeling of extreme tiredness resulting from effortful listening (McGarrigle et al., 2014). The fatigue experienced in everyday situations for individuals with hearing impairment may contribute to higher rates of sick leave than workers without hearing impairment (Kramer, Kapteyn, & Houtgast, 2006). Individuals who are fatigued are also more likely to experience increased levels of psychological distress (Missen, Hollingworth, Eaton, & Crawley, 2012) resulting in a potential recurring cycle of stress and anxiety for individuals with hearing loss (Hétu, Riverin, Lalande, Getty, & St-Cyr, 1988). In order to gain a more comprehensive understanding of the disability associated with hearing loss, many studies have sought to identify reliable markers of listening effort. 'Listening effort' refers to the mental exertion required to attend to, and understand, an auditory message (McGarrigle et al., 2014). It is often assumed that repeated incidences of effortful listening in everyday situations (for individuals with hearing loss) result in the oft-reported experience of mental fatigue (Bess & Hornsby, 2014).

Measures of fatigue can be grouped into three categories: (i) self-report, (ii) behavioural, and (iii) physiological (Anderson Gosselin & Gagne, 2011; McGarrigle et al., 2014). The nature of the relationship between each measure is complex, with physiological and/or behavioural markers of fatigue not always corroborated by self-report measures (Hockey, 2013). Self-report measures (e.g., rating scales and questionnaires) reflect the individual's subjective experience of fatigue. Self-report measures are easy to administer and require little expertise to interpret. However, using self-report measures alone assumes that listeners can reliably perceive (and later, recall) any moment-to-moment changes in fatigue during listening. Further, self-report measures do not provide insight into the underlying physiological mechanisms that underpin the perceptual experience of fatigue (Bess & Hornsby, 2014). Objective markers of fatigue in the form of

behavioural or physiological measures may help in situations where accurate verbal or written responses are not viable (e.g., in individuals with communication disorders), or when fatigue may be underestimated (e.g., in older adults) (Larsby, Hällgren, Lyxell, & Arlinger, 2005).

Longer response times (RTs) over the course of an experiment are believed to represent a behavioural index of fatigue (or 'vigilance') (DeLuca, 2005). During cognitive-motor tasks, RTs have been shown to increase as a function of time on task (Kato, Endo, & Kizuka, 2009; Lorist, Boksem, & Ridderinkhof, 2005; Lorist et al., 2000). These studies suggest that a change in RT reflects a reduction in the allocation of attentional resources following sustained mental processing. The availability of cognitive resources is thought to be essential for maintaining optimal task performance, and the onset of fatigue may give rise to a depletion of such resources (Kato et al., 2009). Only one study to date has investigated fatigue during a listening task. Hornsby (2013) used the dual-task paradigm (see Gosselin and Gagné (2010) for a discussion of this approach) to examine effort and fatigue in hearing-impaired adults listening to speech while wearing hearing aids ('aided') and without hearing aids ('unaided'). Word recognition served as the primary task with both visual RT (i.e., RTs to a visual stimulus) and percentage of words correctly recalled as secondary tasks. Fatigue was indexed behaviourally as the relative increase in visual task RTs over the duration of a one-hour aided and a one-hour unaided trial block. Self-report fatigue was also assessed at the end of the experimental block by asking participants; "*how mentally/physically drained are you right now*?" Relatively less fatigue was exhibited (i.e., visual task RTs showed less of an increase across the one-hour trial block) in the aided versus unaided listening conditions. However, despite a general increase in pre-to-post-task fatigue ratings for all participants, the mean relative change in self-report fatigue ratings did not differ significantly between the aided and unaided conditions (Hornsby, 2013). Therefore, it remains unclear to what extent changes in RT during a listening task are associated with fatigue.

RT can be used as a behavioural marker of fatigue, which may be useful in a clinical setting given its cost-effectiveness and ease of interpretation. However, it does not provide information about the nature of the physiological changes that underpin listening-related effort and fatigue. A better understanding of these physiological changes could provide some insight into the experience of listening-related effort and fatigue (e.g., *when* does fatigue emerge, and *how* is it manifested physiologically?). Pupillometry refers to the method of measuring changes in the size of the eye's pupil, and has been used in the experimental psychology literature to measure attention- and memory-related processes (Beatty, 1982; Kahneman, 1973; Laeng, Sirois, & Gredeback,

2012). The underlying physiological process that governs fluctuations in pupil size is known as the locus coeruleus norepinephrine (LC-NE) system, and has been studied extensively in relation to task performance dynamics and changes in autonomic arousal and alertness (Aston-Jones & Cohen, 2005; Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Hopstaken, van der Linden, Bakker, & Kompier, 2015; Jepma & Nieuwenhuis, 2011). Changes in pupil size have been found to co-vary with changes in the blood oxygen level-dependent (BOLD) response in the locus coeruleus (Murphy, O'Connell, O'Sullivan, Robertson, & Balsters, 2014). Additionally, researchers have observed in animal models that declines in neural function via locus coeruleus projections to prefrontal cortex are thought to underlie cognitive declines associated with neural fatigue (Bellesi, Tononi, Cirelli, & Serra, 2015).

Pupillometry has been used in hearing research to characterise the heightened arousal state associated with increases in listening effort (Kuchinsky et al., 2013; Winn, Edwards, & Litovsky, 2015; Zekveld, Kramer, & Festen, 2010, 2011; Zekveld & Kramer, 2014). On the other hand, mental fatigue may be characterised as a low-arousal state, which is indexed objectively by *decreasing* pupil size over time. Hopstaken et al. (2015) investigated this hypothesis by tracking changes in both baseline and task-evoked pupil size while participants performed the '2-back' task (i.e., participants indicate whether a letter presented on the screen is the same as the letter presented two letters before). Self-report fatigue ratings were also administered following each of the seven trial blocks. Task-evoked pupil size decreased systematically as the experiment progressed, but no differences were found in baseline pupil size. Further, a significant negative correlation (r = -0.33) was found between task-evoked pupil size and self-report fatigue. The authors suggest that LC-NE system-evoked changes in pupil size likely play a role in the effects of mental fatigue. Although the precise neural pathways that connect the locus coeruleus to changes in pupil size are still under debate (Nieuwenhuis, De Geus, & Aston-Jones, 2011), the LC-NE system provides a well-studied psychophysiological framework for understanding mental fatigue (Hopstaken et al., 2015). However, no research studies to date have attempted to examine potential changes in listening-related fatigue using pupillometry.

Pupil size can be monitored continuously during a listening task and can therefore provide useful insight into the onset and extent of mental fatigue, which is likely to change over time. Analysing changes in pupil size over the course of a speech processing task provides a way of measuring listening-related fatigue independently of performance in the behavioural task itself. In hearing research, previous research studies have used pupillometry primarily to assess task-evoked listening effort and have therefore typically adopted

recognition tasks with short duration (word or sentence) stimuli. Little is known about changes in the pupil response over longer timescales (e.g., while listening to longer speech extracts). Neuroimaging studies have demonstrated that sustaining attention engages prefrontal neural systems, which may re-set with the onset of each trial cuing participants' attention to the task (Dosenbach et al., 2006). Previous attempts have been made to characterise the physiological demands of sustained mental processing (Esterman, Noonan, Rosenberg, & DeGutis, 2012; Walter & Porges, 1976). However, to the best of our knowledge no research studies to date have sought to monitor within-trial changes in physiological arousal that occur while listening to speech as it unfolds in the presence of noise. Based on the theory that pupil size reflects dynamic changes in alertness and arousal (Aston-Jones & Cohen, 2005) and the assumption that a reduced state of arousal is a component of fatigue (Hockey, 2013), monitoring within-trial changes in pupil size may reveal fatigue-related changes as a result of sustained listening in noise.

The aim of the present study was to investigate RT and pupillometry as potential markers of listening-related fatigue. A Speech-Picture Verification (SPV) task, used widely in the experimental psychology literature (Zwaan, Stanfield, & Yaxley, 2002), was adapted to include short passages. Two listening conditions were created ('easy' and 'hard') with contrasting SNRs (+15 dB SNR and -8 dB SNR, respectively). For each trial, participants were presented with a narrative speech passage and were then asked to respond to an image indicating (by pressing 'yes' or 'no' on a remote control) whether it corresponded to an object mentioned in the preceding passage. RTs for correctly-responded items were automatically recorded and analysed as a behavioural marker of fatigue. Pupil size was also recorded (while listening to the passage) as a physiological marker of listening-related fatigue. Finally, self-report effort and fatigue scales were administered to assess their sensitivity to changes in listening task demand. We hypothesised:

i. A steeper within-trial decrease in pupil size in the 'hard' versus the 'easy' condition, which is more pronounced in the 2nd versus the 1st half of the listening condition block (i.e., a Condition x block 'Half' interaction); reflecting increased listening-related fatigue during sustained listening.

ii. A steeper RT increase across trials in the 'hard' versus the 'easy' listening condition; reflecting the behavioural marker of listening-related fatigue, as in Hornsby (2013).

iii. Higher self-report effort and fatigue ratings in the 'hard' versus the 'easy' listening condition, reflecting the increased subjective experiences of effort and fatigue in the more challenging SNR, as in Zekveld et al. (2010).

Method

Participants

Twenty-four healthy young adults aged 18 to 30 years took part in this study. This sample size of 24 was calculated based on Zekveld et al. (2010) (d = 0.83), providing a statistical power of 0.8, for a 2-tailed prediction and an alpha level of 0.05. All participants were native English-language speakers who reported: (i) normal or corrected-to-normal visual acuity, and (ii) no language or hearing impairments. Participants were recruited either through flyers posted around the University of Manchester campus or as part of a course credit scheme for Psychology undergraduate students. Participants who did not receive course credit were financially reimbursed for their time. Participants provided informed written consent before participating in the University of Manchester Research Ethics Committee approved study.

Equipment

Participants sat 60 cm away from a 19" flat screen computer monitor, which displayed the visual stimuli. The participant's head was stabilised on a chin-rest secured onto the end of the table. Stimulus presentation was programmed using the SR Research Experiment Builder software (SR Research, Mississauga, ON, Canada). Auditory stimuli were presented through two speakers placed on either side of the computer monitor, at 45° and 315° azimuth.

Pupil size was recorded using an Eyelink 1000, with a sampling rate of 1000 Hz. Pupil size was recorded as an integer number, in arbitrary units, and is related to the number of pixels in the pupil image (i.e., the number of pixels contained within the camera's pupil image). Typical pupil area can range between 100 to 10000 units, with a precision of 1 unit. This corresponds to a resolution of 0.01 mm for a 5 mm pupil diameter. Participants responded to picture targets via a button box interfaced with the Eyelink software. The desktop-mounted eye-tracker was placed on the table in between the participant and the computer monitor (50 cm away

from the participant) at 0° azimuth. The eye-tracker was aligned to the centre of the computer monitor and was as close to the lower edge of the computer monitor as possible to maximise the eye tracking range.

Materials

Pure-tone audiometry. Pure-tone hearing thresholds were measured at the beginning of the experiment to ensure that all participants had hearing thresholds of ≤ 20 dB HL in each ear at 500, 1000, 2000, and 4000 Hz.

Speech-Picture Verification (SPV) task. Each passage contained three sentences and was 45 – 50 words long, with a passage duration of between 13 – 18 seconds. The first sentence of the passage introduced the scene (e.g., '*Bob lives near a beautiful park and loves going for long walks there during Spring*'), the second sentence mentions the target object (e.g., '*This time he decided to bring binoculars to see if he could spot any* ***pigeons*** *in the trees*'), and the final sentence refers back to the target noun using a pronoun referent (e.g., '*Fortunately, he managed to catch a glimpse of one perching in its nest*'). All speech stimuli were recorded by a native British English female speaker. Comprehension difficulty was assessed using the Flesch reading ease scale, a commonly used metric that analyses the statistical properties of a speech extract in contemporary English. The speech material had a mean Flesch reading ease score of 75.9, and a Flesch-Kincaid Grade level of 6.6 (i.e., suitable for individuals aged 10 years or older) (Kincaid, Fishburne Jr, Rogers, & Chissom, 1975).

Two lists of speech-picture pairs (List A and List B) were compiled, using a Latin-squares experimental design (i.e., the same speech materials used in the 'easy' condition in List A were used in the 'hard' condition in List B, etc.). Using a Latin-squares design makes it unlikely that certain experimental details (e.g., lexical frequency, audibility of target object name) can have any systematic effect on the results with respect to the comparisons of interest between easy and hard listening conditions. Participants were randomly allocated to one of the two item lists. Each list contained 46 'yes' responses (i.e., the object depicted in the picture was mentioned in the passage) and 46 'no' responses (i.e., the object depicted in the picture was not mentioned in the passage). Of the 46 'yes' response speech-picture pairs, 40 included the target object (i.e., the object depicted in the image) in the second sentence. For the other six items, the target object was mentioned in either the first or last sentence. These six items were included to encourage participants to attend to the whole

passage rather than just the second sentence. The images used in this experiment were full colour photographs - as used by Engelen, Bouwmeester, de Bruin, and Zwaan (2011). This visual angle subtended by the image extended 20 degrees vertically and 14 degrees horizontally. All images were centred on the screen, with a screen resolution of 1680 x 1050 pixels.

A background noise file consisting of multi-talker babble, taken from the International Collegium of Rehabilitative Audiology (ICRA) CD (Dreschler, Verschuure, Ludvigsen, & Westermann, 2001), was digitally mixed with the speech stimuli to create two listening conditions ('easy' and 'hard'). The SNR for the 'easy' listening condition was + 15 dB. The SNR for the 'hard' listening condition was – 8 dB. These SNRs were chosen based on the performance of normal hearing participants in a pilot study, which revealed that ≥ 90% accuracy was achieved across both conditions, but with participants reporting greater subjective difficulty in the 'hard' condition. Root-mean-square (rms) values were adjusted for each speech and noise file to set the desired SNR, while equalising the overall power between conditions. Overall output level for both listening conditions was fixed at 65 dB (A). Four practice trials were created; two in each listening condition.

Self-report ratings. Participants completed self-report 'effort' and 'fatigue' ratings after each listening condition block. To assess 'effort', we used the following item from the NASA task load index; *'How hard did you have to work to accomplish your level of performance?'* (Hart & Staveland, 1988). This particular scale has previously been used in the literature to assess perceived listening effort (Bologna, Chatterjee, & Dubno, 2013; Mackersie & Cones, 2011). Underneath this instruction was a Likert scale with 21 increments (1 = 'very low', 20 = 'very high'). To assess mental fatigue, we used a subscale of the fatigue Visual Analog Scale (VAS) (Lee, Hicks, & Nino-Murcia, 1991) to include fatigue-related items only (13 in total). For each item, individuals indicated on a scale of 0 – 10 how 'tired', 'drowsy', 'fatigued', etc. they felt (0 = not at all, 10 = extremely). Fatigue scores were calculated as the mean of all 13 items in the questionnaire. This Likert-style version of the scale was chosen as it is easier to administer and score than the original VAS.

Design and Procedure

In order to assess the onset of mental fatigue over time, this experiment implemented a blocked design, i.e., participants completed a block of trials in one condition (e.g., 'easy') followed by a block of trials in the other

condition (e.g., 'hard'). The order of 'easy' and 'hard' trial blocks was counterbalanced so that half of the participants completed the 'easy' block first, while the other half completed the 'hard' block first.

Upon arrival, participants were seated comfortably in a sound-treated booth. Consistent with luminance adjustment procedures reported previously in the literature (Winn et al., 2015; Zekveld et al., 2010), the luminance of the visual field was controlled by adjusting room lighting and screen brightness levels. To avoid floor/ceiling effects in absolute pupil size, each participant's pupil size was recorded in the 'bright' setting (room brightness at 263 lux and screen brightness at 123 cd/m^2), and 'dark' setting (room brightness at 0.28 lux and screen brightness at 0.0019 cd/m^2). Room lighting and screen brightness were then adjusted for each individual to approximate the middle of the 'bright' and 'dark' setting pupil extremes. The corresponding settings were then used for the rest of the experiment.

Eye-tracker calibration was then performed to ensure that eye data could be accurately mapped onto gaze position. Participants were seated in the sound-treated booth and given the following instructions by the experimenter; *'For each trial, you will hear the same female speaker reading a short passage of text in the presence of background noise. Please look straight ahead at the fixation cross while you listen. Shortly after each passage, an image will be presented on the screen. Please indicate by pressing 'yes' or 'no' on the remote control whether or not the object presented was mentioned in the preceding passage. The object can be mentioned anywhere in the passage so it's important that you pay attention to the whole passage. Please respond as quickly and accurately as possible'.* In the event that no response is given after a period of ten seconds, the subsequent trial begins automatically (i.e., participants are given the prompt 'Press any key to continue').

Participants performed four practice trials (two in each condition) before beginning the recorded experiment to familiarise themselves with the task. Participants began each trial by fixating on an asterisk shown in the centre of the screen. Each trial began with the presentation of background noise (multi-talker babble) to prime the listener. After one second of noise-alone presentation, the speech passage began. Each speech passage was also followed by one second of noise-alone presentation to ensure that the listener could anticipate the visual presentation. Immediately following the end of the noise, an image appeared on the screen and the participants responded 'yes' or 'no'. An inter-trial interval of 10 seconds was programmed to ensure that the pupil had returned to a resting (i.e., stable) size following the behavioural response in the preceding trial. Each listening condition block lasted approximately 20 minutes. Following the first and second blocks, participants

were given 'effort' and 'fatigue' self-report scales to complete. Including the instruction period and completion of the practice trials, the experiment lasted approximately 50 minutes in total.

Analyses

Pupillometry.

Preprocessing. Consistent with the previous literature (Kuchinsky et al., 2013; Piquado, Isaacowitz, & Wingfield, 2010; Winn et al., 2015; Zekveld & Kramer, 2014), pupil data were pre-processed to remove noise from the analysis. Any pupil data beyond the first 13 seconds (post speech-onset) were excluded from the analysis. As the shortest duration speech passage was 13-seconds long, this helped to ensure that only speech processing-related changes (and not motor planning or behavioural response artefacts) were included in the analysis. Correct behavioural responses only were included in the analysis given the uncertainty over potential sources of response errors (e.g., lapses in concentration or misperceptions of key words) (Kuchinsky et al., 2014). Behavioural performance accuracy was at ceiling level (95% correct) across both conditions. In total, incorrect responses accounted for only 5.7% of all trials.

Gaze position can influence pupil size recording, causing estimation errors when the pupil is in a rotated position (i.e., looking at the corner of the screen) (Brisson et al., 2013). However, correlation tests revealed no significant relationships between pupil size and gaze location (x, y co-ordinates) in each listening condition (all p-values > 0.05). A rectangular 'interest area' (left, top, right, bottom, and centre screen co-ordinates: 332, 168, 1347, 881, 839) was created in the centre of the visual display, and any erratic fixations (i.e., fixations that fell outside of this perimeter) were removed from the analysis. This limited the potential for any pupil size estimation errors caused by changes in gaze position.

Missing value samples (e.g., due to blinking) were removed from the analysis, and these points were linearly interpolated across using data from previous and subsequent samples. A paired t-test revealed no significant difference in blink rates between listening conditions ($p > 0.05$). Any trials that included > 50% missing data points were removed from the analysis. This resulted in the removal of 125 trials across all participants (5.6% of all trials in the dataset). Like Kuchinsky et al (2013), a within-trial scaling method was used in order to ensure consistent normalisation across both trials and participants. Each data point was divided by the mean of the entire trial for each participant. This method controls for any scaling variability in the data.

Mean pupil size during the one second of noise-alone presentation preceding speech-onset was used as the baseline for each trial. All data were then baseline-corrected (i.e., each trial's baseline value was subtracted from every data point in that trial). Pupil size fluctuations seen in the data therefore reflect relative changes from baseline for each participant and each individual trial. The pre-processed time series data were then averaged providing a mean pupil size sample for every 500 ms of the analysis for each participant in each condition ('easy' and 'hard') and each half of the listening condition block ('1st' and '2nd'). This time window was chosen based on the predicted slow latency change in the pupil slope over time (in the order of 1-2 seconds) (Beatty, 1982).

Growth curve analysis. Growth curve analysis (GCA) is a statistical technique that is considered to be more appropriate for analysing time series data than other traditional approaches, for example time-binned analyses of variance (ANOVAs) (Kuchinsky et al., 2013; Mirman, 2014; Winn et al., 2015). Unlike time-binned ANOVA approaches, GCA: (i) does not require a trade-off between statistical power and temporal resolution by using 'binned' samples, (ii) eliminates potential experimenter bias (e.g., allocating time windows in an arbitrary fashion for analysis), and (iii) provides a method for quantifying individual differences (Mirman, 2014).

The pupil response over time does not always take a linear form (Kuchinsky et al., 2014; Winn et al., 2015). Further, the precise onset and time course of mental fatigue is not yet known (DeLuca, 2005). GCA provides a robust statistical approach for capturing 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, which make each individual polynomial time term (e.g., linear, quadratic, cubic, etc.) independent of one another (Kalénine, Mirman, Middleton, & Buxbaum, 2012). As a result, each polynomial term time captures a distinct functional form. The 'intercept' refers to the overall mean of the pupil response (i.e., 'area under the curve'); the 'linear' term refers to the slope of the pupil response (larger values indicating a steeper slope); the 'quadratic' term refers to the shape of the primary curve inflection point (more positive values indicate a flatter, more linear, shape); and finally the 'cubic' term generally reflects the extent to which there is a secondary inflection point in the pupil curvature (more positive values indicate a more transient, steeply rising and falling, peak pupil response) (Kuchinsky et al., 2014).

GCA was implemented using R (R Development Core Team, 2013). The 'lme4' package was used for mixed-effects modelling (Bates, Maechler, & Bolker, 2013). Fixed effects (i.e., experimental manipulations and

polynomial time terms) and random effects (i.e., error terms) were modelled to predict the pupil response over the course of speech processing. The fixed effect of Condition was a categorical variable with Easy as the reference level. As advised in Barr et al (2013), the initial model included a maximal random effects structure. However, this model did not converge due to over-parameterization. Following the recommendations of Barr (2013) and Mirman (2014), we set out to systematically remove random effects terms that either: (i) did not contribute significantly to model fit based on Likelihood Ratio tests and/or (ii) were of little or no theoretical importance for interpreting the fixed effects. This process continued until the model finally converged. The final (optimal) random effects structure included the highest order interaction term of interest across both subjects and items. Specifically, both subjects and items were allowed to vary in the highest order interaction effect of interest (i.e., Subject x Condition x Half) for each of the polynomial time terms (e.g., intercept, linear, quadratic, and cubic). Subjects were also allowed to vary for each fixed effect polynomial time term (i.e., the overall curvature of their pupil response, *across* Condition and Half, was allowed to vary)[1]. The inclusion of random effect terms that correspond to the effects of interest results in a more conservative estimate of the fixed effects (Barr, 2013; Mirman, 2014). Parameter estimates are reported using maximum likelihood estimation.

In order to assess the impact of our experimental manipulations (Condition and Half) on the fixed effect polynomial time terms, a best-fitting model was determined using backwards elimination model testing. The ultimate analysis used a third-order (i.e., cubic) polynomial model. The intercept through cubic components captures the extent and timing of the primary inflection (i.e., peak response) as well as the steepness of the subsequent slope. As these were the components of the pupil response of theoretical interest, we did not attempt to add higher-order terms (quartic, quintic, etc.) to the model. Like Kuchinsky et al. (2014), our full mixed model included the highest-order polynomial interaction of interest (Cubic x Condition x Half) along with the lower-order polynomials that make up this effect (e.g., Condition x Half, Condition x Cubic, Half x Cubic, Condition, Half, and Cubic). A significant effect was found if removal of the variable of interest only from the model resulted in a significant decrease in the -2 times the change in log likelihood, which is distributed as χ^2 with degrees of freedom equal to the number of parameters added (Mirman, 2014). A lower-order main effect and/or interaction was always included in the model as a fixed effect term if it was subsumed by a higher-order main effect or interaction that significantly improved model fit (Kuchinsky et al., 2014). P values were

[1] The R code used for the final (optimal) model was: *m.fullCubic <- lmer(PupilSize ~ (ot1+ot2+ot3) * Condition * Half + (ot1+ot2+ot3|Subject) + (ot1+ot2+ot3|Item:Condition:Half) + (ot1+ot2+ot3|Subject:Condition:Half), data=BlockStudyGCA, control=lmerControl(optimizer="bobyqa"), REML=FALSE).*

calculated using the *z* distribution as an approximation of the *t* distribution. As degrees of freedom are poorly defined in a hierarchical mixed-effects model, this approach can be used to estimate statistical significance (Mirman, 2014). To clarify significant interactions, submodels were tested in which the effect of Condition on each of the polynomial terms was evaluated separately for the first and the second half of each block.

Response times. RTs were recorded as the time taken between image presentation and correct button press response. Shapiro-Wilk tests revealed that data were not normally distributed ($p < .05$). All RT data were log-transformed to approximate a normal distribution before statistical analysis. A Shapiro-Wilk test revealed that the log-transformed data were normally-distributed ($p > 0.05$). To examine behavioural markers of mental fatigue, the same GCA approach was applied to the log-transformed RT data across experimental trials. Incorrect trial responses were coded as missing values in the analysis. Changes in RT over the course of trials 1 – 46 within each listening condition block were analysed. The same random effects optimisation and model testing procedures were used as is described above. The ultimate analysis used a second-order polynomial, including linear and quadratic components. The linear component was included to capture the predicted steeper increase in RTs over time in the 'hard' versus the 'easy' listening condition. The quadratic component was included to capture a potential U-shaped response curve reflecting an initial improvement in RTs (i.e., practice effect) followed by a rising slope in the latter stages of the block (i.e., fatigue effect). Participant means (i.e., intercepts) were allowed to vary as well the Subject x Condition interaction for intercept, linear, and quadratic components.

Self-report. Given the rank-based nature of self-report ratings, a Wilcoxon signed ranks test was conducted to test whether subjective effort and fatigue ratings differed significantly between listening conditions.

Results

Pupillometry

Table 1 presents the impact of removing each polynomial interaction term on the overall best-fitting cubic model (depicted in Figure 1). There was a significant Condition x Half interaction on the linear term (β = -

0.61, t = -1.97, p = 0.05). In other words, participants showed a steeper negative sloping pupil response for 'hard' versus 'easy' in the 2nd versus the 1st half of the block. Separate submodels (1st and 2nd half of each block) were tested to clarify these interaction effects. In the second half of the block only, there was a significant effect of Condition on the linear term (β = -0.81, t = -3.32, p < 0.001). All other model tests were non-significant (all p-values > 0.05).

Insert Figure 1 here

Insert Table 1 here

Response times

Figure 2 illustrates the observed RT data. Table 2 presents the impact of removing each polynomial term on overall model fit. Overall, model fit improved significantly when adding the effect of 'condition' on the intercept term (χ^2 [1, N = 24] = 6.42, p = 0.01). In other words, there was a significant difference between listening conditions in overall mean RT. Specifically, participants were slower to respond correctly in the 'hard' versus the 'easy' listening condition (β = 0.05, t = 3.72, p < 0.001). However, RTs between listening conditions did not differ in their rate of change (i.e., the linear term) (χ^2 [1, N = 24] = 1.09, p = 0.30), or in terms of the shape of the primary (U-shaped) inflection curve (i.e., the quadratic term) (χ^2 [1, N = 24] = 1.35, p = 0.25).

Insert Figure 2 here

Insert Table 2 here

Self-report

Figure 3 illustrates the observed self-report ratings data. Self-report effort ratings were higher in the 'hard' (median = 14, IQR = 4) versus 'easy' (median = 7, IQR = 9) listening condition. A Wilcoxon signed ranks test revealed that this difference was statistically significant (z = 3.49, p < 0.001, r = 0.71). Though self-report

fatigue ratings were numerically higher in the 'easy' (median = 5, IQR = 2.81) versus the 'hard' (median = 4.89, IQR = 2.98) listening condition, a Wilcoxon signed ranks test revealed that this difference was not statistically significant (z = 1.12, p = 0.27, r = 0.23).

Insert Figure 3 here

Individual differences

Random effect estimates indicate how much an individual's value in each condition differs from the group mean (in that same condition). We were therefore able to use random effect estimates to compute effect sizes for each subject. For example, we subtracted the random effect estimate of how much an individual's pupil size decreased over time (i.e., the linear term) in the 'easy' condition from the 'hard' condition in the 2nd half of each block only (reflecting the 'fatigue' effect). The same method was used for computing pupillometric 'effort' effect size estimates (i.e., subtracting 'easy' from 'hard' on the intercept term for the shorter 2500 ms model), and RT intercept effect sizes (i.e., subtracting 'easy' from 'hard' on the intercept term in the RT random effects). This method provides a way to quantify how individuals vary from the overall group (i.e., fixed effect) pattern (Mirman, 2014), and permits the analysis of how the strength of these effects co-vary and relate to other outcome measures (Kuchinsky et al., 2014). Effect size estimates for self-report effort ratings were computed by subtracting each subject's 'easy' condition effort rating from their 'hard' condition effort rating.

Individuals who showed the largest RT intercept effect size, i.e., individuals who showed a larger effect of listening condition on RTs (relative to the group mean difference), tended to show a smaller difference in self-report effort ratings between listening conditions (r = -0.60, p = 0.002). No other relationships were found between self-report, RT, and pupillometric fatigue (linear) effects (all p-values > 0.05).

Discussion

The present study used a novel listening paradigm to examine whether physiological (pupil size), behavioural (RTs) and self-report changes over time could provide information regarding the fatiguing effect of listening in challenging SNRs. We hypothesised that, in the 'hard' versus the 'easy' listening condition, RTs would show a

steeper increase across trials, while pupil size would show a steeper within-trial decrease in the second versus the first half of the trial block, reflecting the more pronounced onset of listening-related fatigue. While there was an overall difference in RTs between listening conditions, changes in RT over time did not reflect the hypothesised onset of fatigue in this particular task. Self-report effort ratings were higher in the 'hard' versus the 'easy' listening, but there was no difference in self-report fatigue. Changes in pupil size over time reflected the predicted increase in mental fatigue in the 'hard' versus the 'easy' listening condition. The present study extends previous research by using self-report, behavioural, and physiological measures to examine potential changes in listening-related fatigue. Background noise is ubiquitous in everyday environments and individuals are often required to follow continuous dialogue. The current study is the first empirical evidence of a within-trial change in physiological arousal while listening to a speech passage as it unfolds; with a predicted steeper decrease across the experimental block that is consistent with the onset of fatigue.

Pupil size as a sensitive physiological marker of listening-related fatigue

The present study shows physiological changes with: (i) young normal-hearing adults, (ii) a very high level of performance accuracy, (iii) a task duration of < 1 hour, and (iv) no sleep deprivation. Pupillometry may therefore represent a promising tool for the detection of fatigue in more chronic sufferers, such as individuals with hearing loss. Recent evidence in hearing research suggests that pupillometry is sensitive to task-evoked arousal or 'listening effort' (Kuchinsky et al., 2013; Winn et al., 2015; Zekveld et al., 2010, 2011; Zekveld & Kramer, 2014). It is intuitive (and therefore often assumed) that repeated or sustained effortful listening over time will give rise to mental fatigue. Indeed, anecdotal reports of fatigue in hearing-impaired individuals are, at least in part, what motivates our interest in listening effort (Hornsby, 2013; McGarrigle et al., 2014). The present findings provide novel insight into the underlying physiological changes that occur over time during sustained listening. Specifically, the linear interaction term (see Table 1) suggests that pupil size may reflect changes across the duration of an experiment relating to the onset of mental fatigue. Using a listening task with extended speech material permits the analysis of mental fatigue, and may better capture the fatigue associated with everyday listening for individuals with hearing loss compared to more commonly-used speech recognition tasks. In the second half of each trial block, a reduction in pupil size below baseline occurs to some extent in both listening conditions. This is consistent with the idea that, in both conditions, participants experience reduced arousal in the second versus the first half of each trial block. However, the more pronounced pattern of

463 reduced arousal in the 'hard' listening condition suggests that pupil size is modulated by the acoustic demands
464 imposed on the listener.

465 The present findings represent a preliminary step towards a better understanding of the demands
466 experienced by individuals with (and without) hearing loss in challenging listening environments, which
467 frequently require sustained mental processing. However, the extent to which the findings in the current study
468 reflect the everyday experience of fatigue in the clinical population remains unclear. Ultimately, a reliable
469 objective measure of listening-related fatigue could be of value in discriminating relative benefit from different
470 interventions in the clinical population. For example, in cases where two different hearing aids improve speech
471 perception in noise equally well, one hearing aid signal processing strategy may cause relatively less fatigue
472 than the other. Monitoring physiological changes during listening may be a useful tool for assessing the
473 disability associated with listening, especially in individuals who are unable to provide a reliable verbal or
474 behavioural response (e.g., individuals with motor/verbal disabilities). As already discussed, fatigue can have a
475 serious negative impact on an individual's mental health (Bower, 2012; Finsterer & Mahjoub, 2014; Missen et
476 al., 2012). A better understanding of the prevalence of fatigue in hearing loss (a relatively underexplored clinical
477 population) will help to identify and ultimately mitigate this problem. For example, physical exercise and better
478 sleep quality have been shown to decrease levels of fatigue (Rook & Zijlstra, 2006). Targeted interventions such
479 as increasing regular exercise and improving sleep quality may help to reduce the disability associated with
480 hearing loss at the individual level.

481

482 **Response time as a behavioural marker of listening-related fatigue**

483 Based on the findings from Hornsby (2013), we predicted that RTs would show a steeper increase over time in
484 the hard versus the easy listening condition. We found no significant differences in RT change over time
485 between listening conditions. Upon visual inspection of the data (see Figure 2), there does appear to be a
486 decrease (i.e., speeding) of RTs in the middle of the 'hard' listening condition block (between trials 17 – 36).
487 However, GCA did not reveal a significant difference between listening conditions on the quadratic (i.e., U-
488 shaped curve) term. We suggest that this trend in the data may reflect an initial practice effect (i.e., participants
489 become faster at providing responses as they get used to the task) followed by a more delayed effect of mental
490 fatigue (i.e., participants can no longer sustain this level of performance due to fatigue, and therefore RTs
491 increase). However, this interpretation remains speculative given the lack of statistical significance. There was a

492 significant difference between listening conditions in overall mean RT (intercept). Overall, participants
493 responded faster in the 'easy' versus the 'hard' listening condition. This supports previous findings in the
494 literature of slower RTs in more challenging listening conditions during single task listening paradigms
495 (Gatehouse & Gordon, 1990; Houben, van Doorn-Bierman, & Dreschler, 2013). This RT difference likely
496 reflects differences in the effort required when listening under more/less challenging SNRs.

497

498 **Self-report measures of listening effort and listening-related fatigue**

499 Self-report effort and fatigue rating scales were administered to participants following each listening condition
500 block. Although Hornsby (2013) reported behavioural evidence of fatigue in his study, no differences were
501 detected in self-report fatigue using a 1-item scale. We opted to use a more comprehensive 13-item mental
502 fatigue rating scale to detect differences in self-report fatigue following each listening condition block.
503 Participants reported significantly greater 'effort' in the hard versus the easy listening condition. However, no
504 significant difference was detected in self-report fatigue. In other words, although participants found the hard
505 condition more effortful than the easy condition, they reported no differences in fatigue. One possibility is that
506 participants simply did not experience subjective fatigue during this particular task. It is possible that the
507 physiological changes measured using pupillometry did not reach participants' conscious awareness. Indeed, it
508 has been shown that executive processes and awareness of such processes (i.e., the perception of 'effort') during
509 difficult task conditions may have their bases in different neural regions (Naccache et al., 2005). The fatigue
510 (VAS) scale used in this study may therefore not sensitively detect *listening*-related fatigue reported in
511 individuals with hearing loss. Although a more comprehensive (13-item) fatigue assessment scale was used, it
512 may be the case that these questions (e.g., '*please indicate how tired/sleepy/drowsy/fatigued you are right now'*)
513 do not capture the kind of communication-related fatigue experienced following effortful listening. Another
514 potential limitation of these subjective reports relate to their timing. Administering one fatigue scale at the end
515 of an entire trial block limits our ability to track potential changes in subjective fatigue over time. Perhaps
516 fatigue assessment scales at more regular intervals, as in Hopstaken et al. (2015) would enable a more sensitive
517 recording of change over time (as revealed in the pupillometry data).

518

519 **Individual differences**

Analysis of individual differences revealed that participants who showed a larger mean RT difference between 'hard' and 'easy' listening conditions (relative to the overall group effect) tended to report a smaller difference in subjective effort between these listening conditions. This finding appears counter-intuitive as it is not consistent with findings in the literature that slower RTs are indicative of increased mental effort (Anderson Gosselin & Gagne, 2011; Fraser, Gagne, Alepins, & Dubois, 2010; Gatehouse & Gordon, 1990; Hornsby, 2013; Houben et al., 2013; Sarampalis, Kalluri, Edwards, & Hafter, 2009). In fact, the opposite pattern is found in the present study; individuals who found the 'hard' condition subjectively more effortful than the 'easy' condition, generally showed a smaller RT difference between conditions. There may therefore be variation between individuals in the extent to which 'effort' is allocated to maintain performance in a behavioural task. Indeed, these data suggests that the increased perceived allocation of 'effort' may not always be accompanied by an increase in behavioural RTs. One potential reason for this unexpected finding may relate to the nature of the self-report effort question posed; *"How hard did you have to work to accomplish your level of performance?"* It is possible that some participants interpreted 'level of performance' as pertaining to the speed of their responses (as opposed to their performance accuracy). In other words, participants who showed less of a RT discrepancy between conditions (i.e., RTs were maintained at a relatively fast speed in the 'hard' condition) did so at the expense of increased levels of perceived effort. Another potential explanation is that some individuals may simply have tried harder (resulting in higher effort ratings), and consequently had faster reaction times in the hard condition (i.e., they were less different from the easy condition). This unexpected relationship highlights the importance of: (i) using an independent (in this case, physiological) measure in research studies, and (ii) being careful to ensure that self-report questions directly address the construct of interest (i.e., the 'effort' required to maintain *accurate* performance).

Limitations and Future research

The high levels of performance accuracy observed in the current study may lead one to question whether the physiological changes observed are a reflection of participant boredom, which may present itself in a similar manner to fatigue (e.g., reduced arousal). However, if participants were experiencing boredom, one would expect to observe either of the following two effects: (i) a similar pattern of decreased physiological arousal across both listening conditions (i.e., not specific to the 'hard' condition), or (ii) a more pronounced reduction in physiological arousal in the less challenging (and consequently more 'boring') condition. The observed decrease

in physiological arousal specifically in the more challenging ('hard') condition makes this interpretation unlikely and lends support to the idea that participants are experiencing a reduction in physiological arousal as a result of sustained listening demands.

Given the high level of performance accuracy in both listening conditions during both pilot testing and the actual experimental data, we did not test the hypothesis that there would be a difference in the change in performance accuracy across trials. However, it is known that decrements in performance accuracy may occur as a result of fatigue during a cognitive task (Hockey, 2013). We found no significant main effect of block half ($F = 1.255$, $p = 0.27$), and no interaction between listening condition and block half ($F = 0.489$, $p = 0.49$). In other words, we found no decline in performance accuracy as a function of listening condition. It therefore appears that the physiological change observed in the more challenging listening condition occurred at the expense of maintaining correct performance accuracy.

The one second of noise-alone presentation preceding speech onset was used as a baseline in this study in order to delineate the pupil response during listening to speech in noise from any physiological response to the presence of noise alone. However, it is possible that the contrasting pupil response slopes observed in the data may have been influenced by differences between conditions in pupil size in response to the noise-alone presentation used as a baseline. In order to address this concern, we analysed the impact of listening condition and block half on pupil size during the one second of noise-alone presentation. We found no significant main effect of listening condition, $F(1, 23) = 0.582$, $p = 0.45$, nor any interaction between listening condition and block half, $F(1, 23) = 1.994$, $p = 0.17$. This suggests that the choice of baseline did not influence the pupil response difference that we found between listening conditions.

Given the large SNR difference between listening conditions used in the present study (23 dB), future research could explore the extent to which pupillometric markers of mental fatigue are sensitive to more subtle differences in SNR as well as other factors (e.g., reverberation, accented speech). Certain cognitive factors (e.g., working memory span) are believed to predict hearing aid success and successful speech understanding in noise (Akeroyd, 2008; Rönnberg, Rudner, Lunner, & Zekveld, 2010). Further research into the cognitive predictors of listening-related fatigue may also shed light on possible underlying cognitive factors that influence the extent to which an individual becomes mentally fatigued due to listening demands. Hearing loss is a predictor of stress-related absence from occupational work (Kramer et al., 2006). However, it is unclear to what extent mental

fatigue in chronic hearing loss may impact everyday functioning. More research is needed to investigate the consequences of increased listening-related fatigue in individuals with hearing loss.

The current study set out to measure within-trial physiological changes while listening to a speech passage as it unfolds. Previous research has identified a reduction in the task-evoked pupil response during a visual working-memory task as a potential marker of fatigue (Hopstaken, et al., 2015). The extent to which changes across trials in the task-evoked pupil response during a listening task reveal information about fatigue remains an important question for future research. The listening task used in the present study involved speech passages between 13 – 18 seconds duration. Although the location of the target word within a passage was periodically varied, it was most frequently mentioned around the middle of the passage (i.e., the second sentence). The question over whether or not there is a change over time in processing load (i.e., the task-evoked pupil response) could be empirically tested by using shorter duration auditory stimuli (e.g., word or sentence) that are time-locked to the canonical task-evoked pupil response. This response typically occurs within two seconds of stimulus onset.

Conclusions

The present study provides evidence of a potential physiological marker of mental fatigue during a listening task. An objective measure of listening-related fatigue could help to provide a more comprehensive understanding of the disability associated with hearing loss. Physiological measures may serve an important role in detecting the potentially damaging effects of fatigue. A sensitive and reliable measure of listening-related fatigue may also be used in the audiology clinic as a measure of benefit from hearing devices and other types of intervention (e.g., auditory training). A better understanding of the prevalence of fatigue associated with hearing loss will ultimately help to remediate this problem by identifying appropriate intervention treatments for individuals.

Acknowledgements

The authors would like to thank the Castang Foundation for their generous funding towards this work. We would also like to thank Kathryn Hopkins, Keith Wilbraham, and Richard Baker for their help with the

recording of speech stimuli and their technical assistance. Finally, thank you to Marcus Johnson at SR Research for his technical support with both the eye-tracker and the experimental design.

References

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47*(S2), 53-71. doi: 10.1080/14992020802301142

Anderson Gosselin, P., & Gagne, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language and Hearing Research, 54*(3), 944. doi:10.1044/1092-4388(2010/10-0069)

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. doi: 10.1146/annurev.neuro.28.061604.135709

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*, 328. doi: 10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255-278. doi:10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., & Bolker, B. (2013). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0. 2012. *URL: http://CRAN. R-project. org/package= lme4*.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276-292. doi:10.1037/0033-2909.91.2.276

Bellesi, M., Tononi, G., Cirelli, C., & Serra, P. A. (2015). Region-specific dissociation between cortical noradrenaline levels and the sleep/wake cycle. *Sleep, 39*(1), 143-154. doi: 10.5665/sleep.5336.

Bess, F. H., & Hornsby, B. W. (2014). Commentary: Listening can be exhausting—Fatigue in children and adults with hearing loss. *Ear and Hearing, 35*(6), 592-599. doi: 10.1097/AUD.0000000000000099

Bologna, W. J., Chatterjee, M., & Dubno, J. R. (2013). Perceived listening effort for a tonal task with contralateral competing signals. *The Journal of the Acoustical Society of America, 134*(4), EL352-EL358. doi: 10.1121/1.4820808.

633 Bower, J. E. (2012). Fatigue, brain, behavior, and immunity: summary of the 2012 Named Series on fatigue.
634 *Brain, Behavior, and Immunity, 26*(8), 1220-1223. doi:10.1016/j.bbi.2012.08.009
635 Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter
636 measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research*
637 *Methods, 45*(4), 1322-1331. doi: 10.3758/s13428-013-0327-0
638 DeLuca, J. (2005). *Fatigue as a window to the brain*: Cambridge: MIT Press.
639 Dosenbach, N. U., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., . . . Petersen, S.
640 E. (2006). A core system for the implementation of task sets. *Neuron, 50*(5), 799-812. doi:
641 10.1016/j.neuron.2006.04.031
642 Dreschler, W. A., Verschuure, H., Ludvigsen, C., & Westermann, S. (2001). ICRA Noises: Artificial Noise
643 Signals with Speech-like Spectral and Temporal Properties for Hearing Instrument Assessment: Ruidos
644 ICRA. *International Journal of Audiology, 40*(3), 148-157. doi: 10.3109/00206090109073110
645 Engelen, J. A., Bouwmeester, S., de Bruin, A. B., & Zwaan, R. A. (2011). Perceptual simulation in developing
646 language comprehension. *Journal of Experimental Child Psychology, 110*(4), 659-675.
647 doi:10.1016/j.jecp.2011.06.009
648 Esterman, M., Noonan, S. K., Rosenberg, M., & DeGutis, J. (2012). In the zone or zoning out? Tracking
649 behavioral and neural fluctuations during sustained attention. *Cerebral Cortex*, *23,* 2712-2723. doi:
650 10.1093/cercor/bhs261
651 Finsterer, J., & Mahjoub, S. Z. (2014). Fatigue in Healthy and Diseased Individuals. *American Journal of*
652 *Hospice & Palliative Medicine, 31*(5), 562-575. doi: 10.1177/1049909113494748
653 Fraser, S., Gagne, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech
654 in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech,*
655 *Language and Hearing Research, 53*(1), 18-33. doi:10.1044/1092-4388(2009/08-0140)
656 Gatehouse, S., & Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification.
657 *British Journal of Audiology, 24*(1), 63-68. doi: 10.3109/03005369009077843
658 Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control
659 state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, &*
660 *Behavioral Neuroscience, 10*(2), 252-269. doi: 10.3758/CABN.10.2.252
661 Gosselin, P. A., & Gagné, J. P. (2010). Use of a Dual-Task Paradigm to Measure Listening Effort. *Inscription*
662 *au Répertoire, 34*(1), 43-51.

663 Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX task load index results of empirical and
664 theoretical research, *Advances in Psychology, 51, 139-183. doi:10.1016/S0166-4115(08)62386-9*
665 Hétu, R., Riverin, L., Lalande, N., Getty, L., & St-Cyr, C. (1988). Qualitative analysis of the handicap
666 associated with occupational hearing loss. *British Journal of Audiology, 22*(4), 251-264. doi:
667 10.3109/03005368809076462
668 Hockey, R. (2013). *The psychology of fatigue: work, effort and control*: Cambridge University Press.
669 Hopstaken, J. F., van der Linden, D., Bakker, A. B., & Kompier, M. A. (2015). The Window of My Eyes: Task
670 Disengagement and Mental Fatigue Covary with Pupil Dynamics. *Biological Psychology, 110,* 100-
671 106. doi:10.1016/j.biopsycho.2015.06.013
672 Hornsby, B. W. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with
673 sustained speech processing demands. *Ear and Hearing, 34*(5), 523-534. doi:
674 10.1097/AUD.0b013e31828003d8
675 Houben, R., van Doorn-Bierman, M., & Dreschler, W. A. (2013). Using response time to speech as a measure
676 for listening effort. *International Journal of Audiology, 52*(11), 753-761. doi:
677 10.3109/14992027.2013.832415
678 Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off:
679 evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience, 23*(7), 1587-1596.
680 doi:10.1162/jocn.2010.21548
681 Kahneman, D. (1973). *Attention and effort*: New Jersey: Prentice-Hall Inc.
682 Kalénine, S., Mirman, D., Middleton, E. L., & Buxbaum, L. J. (2012). Temporal dynamics of activation of
683 thematic and functional knowledge during conceptual processing of manipulable artifacts. *Journal of*
684 *Experimental Psychology: Learning, Memory, and Cognition, 38,* 1274-1295. doi: 10.1037/a0027626
685 Kato, Y., Endo, H., & Kizuka, T. (2009). Mental fatigue and impaired response processes: event-related brain
686 potentials in a Go/NoGo task. *International Journal of Psychophysiology, 72*(2), 204-211. doi:
687 10.1016/j.ijpsycho.2008.12.008
688 Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability*
689 *formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted*
690 *personnel.* Millington, TN: Naval research branch report.

691 Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing
692 and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *International*
693 *Journal of Audiology, 45*(9), 503-512. doi: 10.1080/14992020600754583
694 Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2014). Speech-
695 perception training for older adults with hearing loss impacts word recognition and effort.
696 *Psychophysiology, 51*(10), 1046-1057. doi: 10.1111/psyp.12242
697 Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013).
698 Pupil size varies with word listening and response selection difficulty in older adults with hearing loss.
699 *Psychophysiology, 50*(1), 23-34. doi: 10.1111/j.1469-8986.2012.01477.x.
700 Laeng, B., Sirois, S., & Gredeback, G. (2012). Pupillometry: A Window to the Preconscious? *Perspectives on*
701 *Psychological Science, 7*(1), 18-27. doi: 10.1177/1745691611427305
702 Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in
703 speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired
704 subjects. *International Journal of Audiology, 44*(3), 131-143. doi: 10.1080/14992020500057244
705 Lee, K. A., Hicks, G., & Nino-Murcia, G. (1991). Validity and reliability of a scale to assess fatigue. *Psychiatry*
706 *Research, 36*(3), 291-298. doi:10.1016/0165-1781(91)90027-M
707 Lorist, M. M., Boksem, M. A., & Ridderinkhof, K. R. (2005). Impaired cognitive control and reduced cingulate
708 activity during mental fatigue. *Cognitive Brain Research, 24*(2), 199-205. doi:
709 10.1016/j.cogbrainres.2005.01.018
710 Lorist, M. M., Klein, M., Nieuwenhuis, S., Jong, R., Mulder, G., & Meijman, T. F. (2000). Mental fatigue and
711 task control: planning and preparation. *Psychophysiology, 37*(5), 614-625. doi: 10.1111/1469-
712 8986.3750614
713 Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indices of listening effort in a
714 competing-talker task. *Journal of the American Academy of Audiology, 22*(2), 113-122. doi:
715 10.3766/jaaa.22.2.6
716 McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014).
717 Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology
718 Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology, 53*(7),
719 433-440. doi: 10.3109/14992027.2014.890296
720 Mirman, D. (2014). *Growth curve analysis and visualization using R*: New York: CRC Press.

721 Missen, A., Hollingworth, W., Eaton, N., & Crawley, E. (2012). The financial and psychological impacts on
722 mothers of children with chronic fatigue syndrome (CFS/ME). *Child Care Health and Development,*
723 *38*(4), 505-512. doi: 10.1111/j.1365-2214.2011.01298.x
724 Murphy, P. R., O'Connell, R. G., O'Sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter
725 covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping, 35*(8), 4140-4154. doi:
726 10.1002/hbm.22466
727 Naccache, L., Dehaene, S., Cohen, L., Habert, M.-O., Guichart-Gomez, E., Galanaud, D., & Willer, J.-C.
728 (2005). Effortless control: executive attention and conscious feeling of mental effort are dissociable.
729 *Neuropsychologia, 43*(9), 1318-1328. doi: 10.1016/j.neuropsychologia.2004.11.024
730 Nieuwenhuis, S., De Geus, E. J., & Aston-Jones, G. (2011). The anatomical and functional relationship between
731 the P3 and autonomic components of the orienting response. *Psychophysiology, 48*(2), 162-175. doi:
732 10.1111/j.1469-8986.2010.01057.x
733 Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger
734 and older adults. *Psychophysiology, 47*(3), 560-569. doi: 10.1111/j.1469-8986.2009.00947.x
735 Rönnberg, J., Rudner, M., Lunner, T., & Zekveld, A. (2010). When cognition kicks in: Working memory and
736 speech understanding in noise. *Noise and Health, 12*, 263-269. doi: 10.4103/1463-1741.70505
737 Rook, J. W., & Zijlstra, F. R. H. (2006). The contribution of various types of activities to recovery. *European*
738 *Journal of Work and Organizational Psychology, 15*(2), 218-240. doi: 10.1080/13594320500513962
739 Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of
740 background noise and noise reduction. *Journal of Speech, Language and Hearing Research, 52*(5),
741 1230-1240. doi:10.1044/1092-4388(2009/08-0111)
742 Walter, G. F., & Porges, S. W. (1976). Heart rate and respiratory responses as a function of task difficulty: The
743 use of discriminant analysis in the selection of psychologically sensitive physiological responses.
744 *Psychophysiology, 13*(6), 563-571. doi: 10.1111/j.1469-8986.1976.tb00882.x
745 Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The Impact of Auditory Spectral Resolution on
746 Listening Effort Revealed by Pupil Dilation. *Ear and Hearing*, *36*, e153-e165. doi:
747 10.1097/AUD.0000000000000145
748 Zekveld, A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The
749 influence of sentence intelligibility. *Ear and Hearing, 31*(4), 480-490. doi:
750 10.1097/AUD.0b013e3181d4f251

751 Zekveld, A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: the
752 influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing, 32*(4), 498-510.
753 doi: 10.1097/AUD.0b013e31820512bb
754 Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions:
755 Insights from pupillometry. *Psychophysiology, 51*(3), 277-284. doi: 10.1111/psyp.12151
756 Zwaan, R. A., Stanfield, R. A., & Yaxley, R. H. (2002). Language comprehenders mentally represent the shapes
757 of objects. *Psychological Science, 13*(2), 168-171. doi: 10.1111/1467-9280.00430

758

759

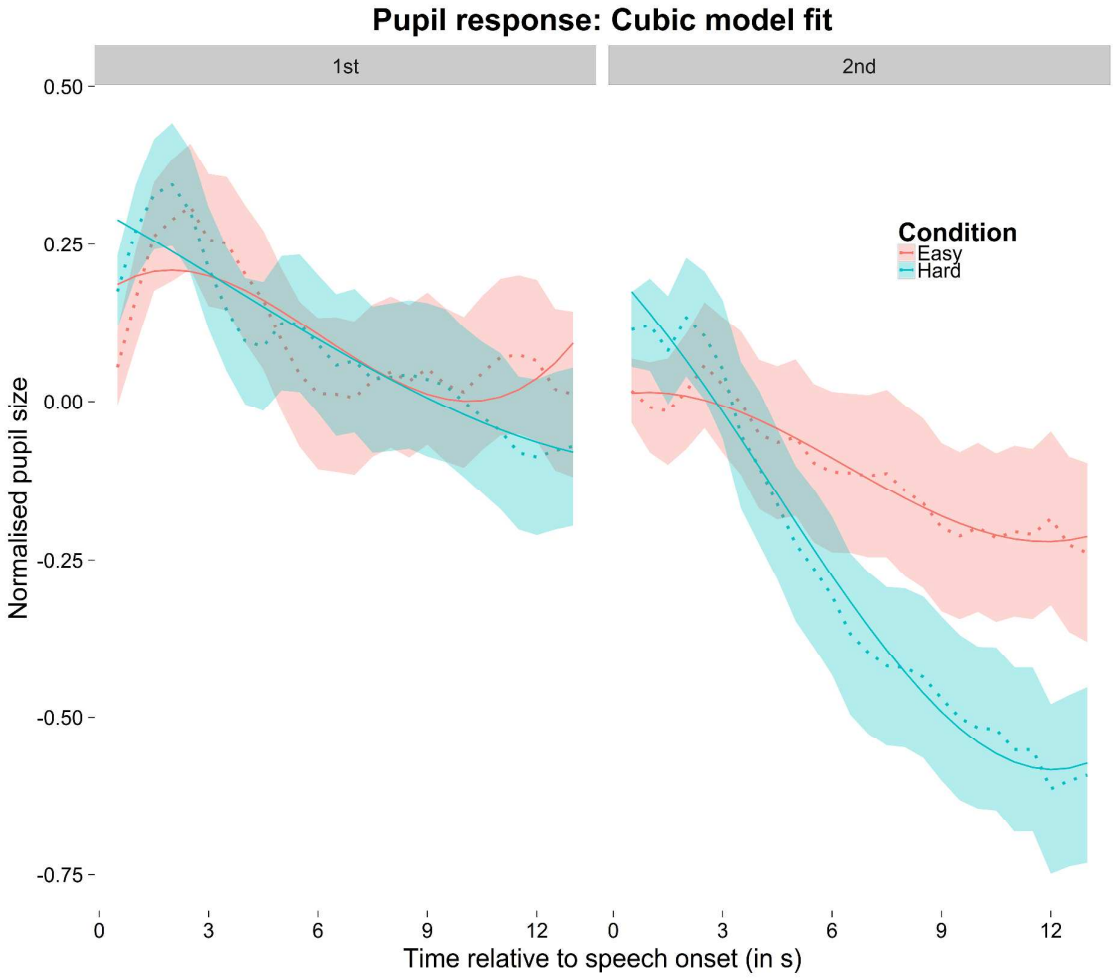

Figure 1. Normalised mean pupil size data (dotted lines, with 95% CI band) overlaid with the GCA cubic model fit (solid lines). These data are plotted as a function of time (second) from speech onset. Left panel, across-subject mean pupil response in 1st half of each block; right panel, across-subject mean pupil response in 2nd half of each block. The plot represents the pupil response during listening only, as the shortest speech passage was 13-seconds long.

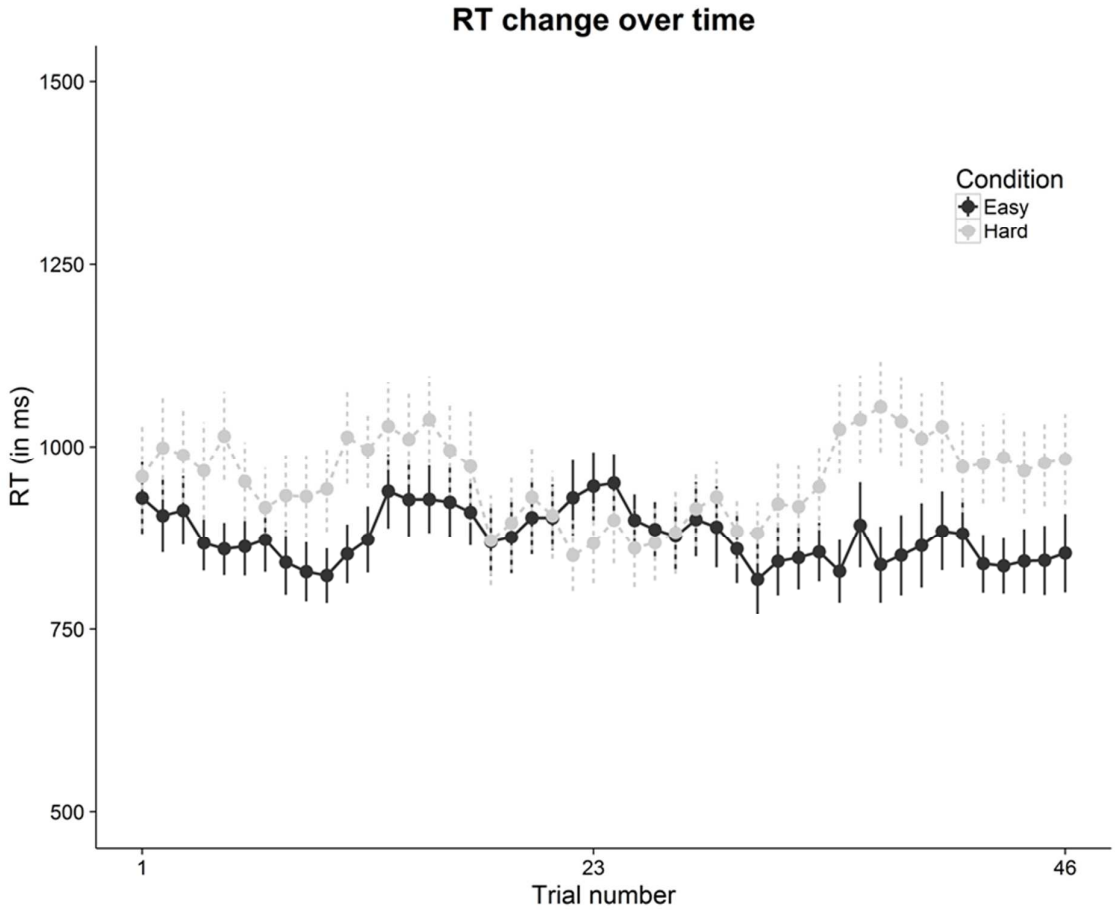

Figure 2. Across-subject mean RTs (for correct responses only) as a function of trial number across each listening condition block (circles with ± SE vertical lines).

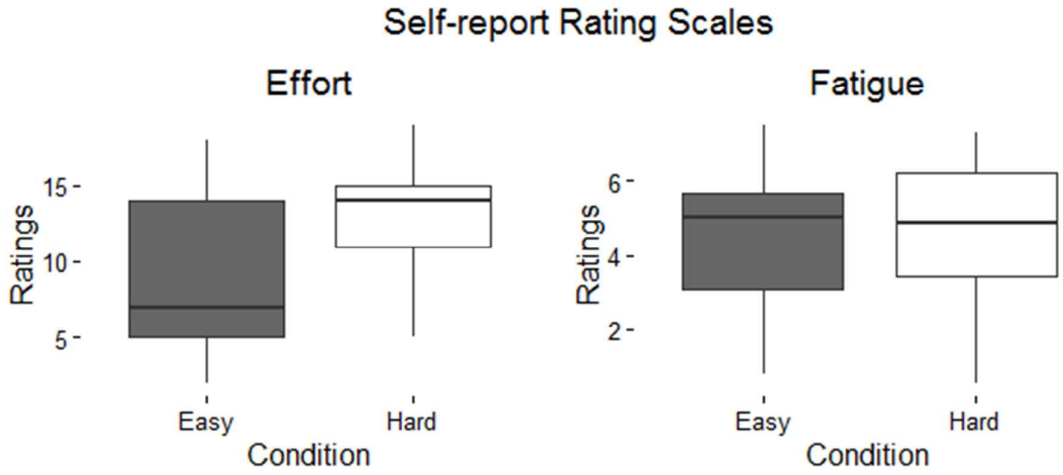

Figure 3. Boxplots for self-report effort and fatigue ratings administered immediately after each listening condition block.

Table 1. Model comparisons to examine changes in the pupil response over time between listening conditions

	Model testing	
Condition x Half x Polynomial	χ^2	*p*
Intercept	0.61	0.44
Linear	**4.23**	**0.04**
Quadratic	0.26	0.61
Cubic	1.36	0.24

Note. Each row presents the extent to which model fit is significantly worse after the removal of a specific polynomial interaction term from the full (cubic) model. Significant effects ($p < 0.05$) are in bold.

Table 2. Model comparisons to examine changes in RT over time between listening conditions

Condition x Polynomial	Model testing	
	χ^2	*p*
Intercept	**6.42**	**0.01**
Linear	1.09	0.30
Quadratic	1.35	0.25

Note. Each row presents the extent to which model fit is significantly worse after the removal of a specific polynomial interaction term from the full (quadratic) model. Significant effects ($p < 0.05$) are in bold.