
Article:
McGarrigle, Ronan Arthur, Rakusen, Lyndon and Mattys, Sven orcid.org/0000-0001-6542-585X (2020) Effortful listening under the microscope : examining relations between pupillometric and subjective markers of effort and tiredness from listening. Psychophysiology. e13703. ISSN 0048-5772

https://doi.org/10.1111/psyp.13703

Received: 18 March 2020 | Revised: 14 September 2020 | Accepted: 16 September 2020

DOI: 10.1111/psyp.13703

ORIGINAL ARTICLE

PSYCHOPHYSIOLOGY SPR WILEY

Effortful listening under the microscope: Examining relations between pupillometric and subjective markers of effort and tiredness from listening

Ronan McGarrigle | **Lyndon Rakusen** | **Sven Mattys**

Department of Psychology, University of York, York, UK

Correspondence
Ronan McGarrigle, Department of Psychology, University of York, York YO10 5DD, UK.
Email: ronanomcg@gmail.com

Funding information
Economic and Social Research Council, Grant/Award Number: ES/R003572/1

Abstract

Effort during listening is commonly measured using the task-evoked pupil response (TEPR); a pupillometric marker of physiological arousal. However, studies to date report no association between TEPR and perceived effort. One possible reason for this is the way in which self-report effort measures are typically administered, namely as a single data point collected at the end of a testing session. Another possible reason is that TEPR might relate more closely to the experience of tiredness from listening than to effort per se. To examine these possibilities, we conducted two preregistered experiments that recorded subjective ratings of effort and tiredness from listening at multiple time points and examined their covariance with TEPR over the course of listening tasks varying in levels of acoustic and attentional demand. In both experiments, we showed a within-subject association between TEPR and tiredness from listening, but no association between TEPR and effort. The data also suggest that the effect of task difficulty on the experience of tiredness from listening may go undetected using the traditional approach of collecting a single data point at the end of a listening block. Finally, this study demonstrates the utility of a novel correlation analysis technique ("rmcorr"), which can be used to overcome statistical power constraints commonly found in the literature. Teasing apart the subjective and physiological mechanisms that underpin effortful listening is a crucial step toward addressing these difficulties in older and/or hearing-impaired individuals.

KEYWORDS

Fatigue, Listening efforListening effort, Pupillometry, Rmcorr, Speech perception

1 | INTRODUCTION

Understanding speech in everyday environments is fraught with challenges arising from a variety of sources, including the level and/or type of interfering acoustic signals present as well as the sensory-cognitive profile of the listener (Mattys et al., 2012). This is because successful speech understanding in adverse conditions relies upon both the fidelity of the acoustic signal that impinges our senses and the "top-down" cognitive mechanisms and linguistic knowledge that help to make sense of the incoming signal (McClelland & Elman, 1986; Pisoni, 1985; Rönnberg et al., 2008). Speech understanding difficulty is exacerbated when to-be-ignored background sounds contain

meaningful information; a phenomenon often referred to as "informational" masking (Kidd et al., 2008). Indeed, the presence of a competing talker during listening can lead to poorer speech understanding ability (Agus et al., 2009) as well as a more negative perception of speech understanding performance (Agus et al., 2009).

However, speech understanding in the presence of a competing talker does not only incur costs in terms of intelligibility. There is growing interest in uncovering ways to measure not just an individual's ability to recognize speech, but also the cognitive effort required to achieve this goal. This is often referred to as "listening effort"; defined recently as "the deliberate allocation ofresources to overcome obstacles in goal pursuit when carrying out a (listening) task." (Pichora-Fuller et al., 2016). In particular, it is believed that a better understanding of mental effort allocation will allow a more comprehensive picture of hearing impairment (McGarrigle et al., 2014). Effortful listening is also a common experience for individuals listening to speech in a second (i.e., nonnative) language or in an unfamiliar accent (Borghini & Hazan, 2018; McLaughlin & Van Engen, 2020). Repeated or sustained episodes of effortful listening may lead to an exacerbated sense of tiredness or fatigue.[1] The types of measures commonly used to extract information relating to effortful listening vary from subjective measures (e.g., self-report questionnaires) to behavioral (e.g., response times) and physiological (e.g., measuring brain activity either directly or indirectly).

Self-report measures of listening effort provide important and ecologically valid insights about the subjective experience of effortful listening. However, they may be prone to bias (Moore & Picou, 2018) and provide limited information about the underlying physiological mechanisms involved. In recent years, there has been a spike in the number of studies using physiological measures to monitor listening effort (see Francis & Love, 2020 for a summary). A commonly used physiological marker of listening effort can be determined using pupillometry, an eye-tracking technique. Fluctuations in the size of the eye's pupil reflect not just adaptive changes to environmental light (e.g., the "light reflex"), but also cognitive-evoked changes that can be traced to changing activity patterns in the brain stem's locus coeruleus (Mathôt, 2018). The locus coeruleus sends and receives projections to and from the cortex, and is thought to govern moment-to-moment changes in our states of attention and arousal (Aston-Jones & Cohen, 2005). In the context of hearing research, the task-evoked pupil response (TEPR) has widely been shown to be sensitive to the increased demands of listening in suboptimal acoustic conditions (see Zekveld et al., 2018 for a review), and is therefore, thought to reflect the effort required to achieve speech recognition under degraded listening conditions. However, while TEPR has consistently been shown to be sensitive to task demand, it has not yet been found to correlate with subjective reports of effort (Koelewijn et al., 2012; Strand et al., 2018) or fatigue (McGarrigle et al., 2017; Wang et al., 2018) during a listening task.

Koelewijn et al. (2012) examined the effect of speech reception threshold (SRT; 50%, 84%) and masker type (single-talker, stationary noise, and fluctuating noise) on TEPR and subjective effort in a group of normal-hearing adults. SRTs were calculated for each participant by adjusting the level of the target speaker relative to the level of the masker until a prespecified performance level was determined. In this case, SRT 50% represents the more-challenging listening condition (i.e., performance ~ 50% correct), and SRT 84% represents the less-challenging condition (i.e., performance ~ 84% correct). Subjective effort ratings were administered at the end of each condition block on a continuous scale. Overall, TEPRs were larger and self-reported effort ratings higher in the single-talker masker condition relative to the stationary and fluctuating noise masker conditions. Further, TEPRs and effort ratings were also sensitive to listening demand; the SRT50% showed larger TEPRs and higher effort ratings compared to the SRT84% condition. However, correlation analyses revealed no significant associations between subjective effort and TEPR; a finding that is consistently reported in the literature (McGarrigle et al., 2014; Pichora-Fuller et al., 2016; Strand et al., 2018).

While evidence for an association between subjective and physiological markers of effort is clearly lacking, there are also conflicting views on the extent to which (or the circumstances under which) the experience of "effort," as measured in laboratory settings, may lead to the experience of "fatigue" from listening (Francis & Love, 2020; Hornsby et al., 2016; McGarrigle et al., 2014). Fatigue is a multifaceted construct that has been operationalized using subjective, behavioral, and physiological markers (Hockey, 2013). However, frequent anecdotal reports of tiredness and fatigue from listening in individuals with a hearing loss have sparked an interest in the subjective manifestation of fatigue (Alhanbali et al., 2017; Holman et al., 2019; Hornsby & Kipp, 2016; Hornsby et al., 2016). Importantly, like effort, the experience of tiredness from listening is not currently tractable based on standard speech understanding assessment procedures alone (e.g., speech recognition performance).

The number of studies investigating listening-related fatigue using subjective and physiological measures have increased in recent years (McGarrigle et al., 2017; Moore et al., 2017; Wang et al., 2018). Wang et al. (2018) found a significant negative correlation between TEPR and reports of daily life fatigue; individuals who reported more daily fatigue had smaller peak TEPRs. However, self-report measures of effort or fatigue were not administered during

[1]We use the terms "tiredness" and "fatigue" interchangeably. "Fatigue" from listening is the terminology most commonly found in the literature, but the scale administered in the present study refers specifically to "tiredness" from listening.

or after the experimental task. McGarrigle et al. (2017) examined the effect of signal-to-noise ratio (SNR) on TEPR and subjective reports of effort and fatigue during a sustained listening task. Following the early task-evoked peak response, pupil size showed a more pronounced downward linear slope during trials in the latter stages of the experiment and in particular for "hard" versus "easy" SNR conditions, suggesting a reduction in the ability to sustain attention and arousal during the more demanding listening condition. Self-reported effort (but not fatigue) varied as a function of SNR. However, no associations were found between TEPR and subjective reports of either effort or fatigue. Likewise, in Moore et al. (2017), participants performed a sustained auditory processing task with a fixed task-demand level while their EEG activity was recorded. Overall, participants reported increased fatigue following the auditory processing task, suggesting that sustained auditory processing can elicit mental fatigue. However, as with previous studies, no relationship was found when assessing the association between subjective and physiological markers of fatigue.

In the studies described above, single self-report evaluations of effort (Koelewijn et al., 2012) and fatigue (McGarrigle et al., 2017; Moore et al., 2017) were collected immediately after each condition of interest, an approach that is fairly standard in the literature (Alhanbali et al., 2017; Dimitrijevic et al., 2019; Rovetti et al., 2019; Strand et al., 2018). However, it is possible that the subjective perceptions of effort and tiredness from listening may fluctuate over the course of a listening experience. While this is often taken into account when recording physiological activity like electroencephalography (EEG) and pupillometry by recording at the level of individual trials, concomitant changes in subjective experiences are rarely examined with a similar level of sensitivity. As a result, the lack of an association between subjective and physiological measures may at least partly reflect inherent differences in the precision with which they are measured. Further, in studies that manipulate task difficulty (e.g., SNR), subjective judgments of effort or fatigue are likely influenced by the conscious perception of a change in task demand or performance. In other words, a listener who becomes subjectively aware of either a change in task demand and/or a change in their own task performance, will likely use these more intuitive judgments to inform their effort or tiredness ratings. Subjective judgments of effort, in particular, are shown to be inversely correlated with performance evaluation (Moore & Picou, 2018).

Systematic examinations of the relationship between domain-general mental fatigue and TEPR can be found in the wider literature. Hopstaken et al. (2015) examined associations between subjective mental fatigue and TEPR over the course of a visual working memory ("2-back") task. Subjective fatigue scales were administered on seven consecutive occasions over the course of the task. The authors found that TEPRs became smaller with higher ratings of mental fatigue, suggesting that when subjective and physiological measures are recorded and analyzed over more frequent time intervals, TEPR appears to be related to the experience of fatigue. In a separate study, Gergelyfi et al. (2015) examined associations between subjective fatigue and a host of physiological measures (including EEG, skin conductance, and pupillometry) while participants performed Sudoku puzzles. In contrast to Hopstaken et al. (2015), no association was found between subjective reports of mental fatigue and TEPRs. These conflicting results suggest that the relationship between TEPRs and subjective fatigue is more complex than initially assumed. Further, as only mental fatigue (and not effort) was examined, it is difficult to ascertain whether TEPR is related more to the experience of effort or to fatigue.

To summarize, despite a rapidly growing literature highlighting the use of pupillometry as an objective measure of listening effort (cf. Zekveld et al., 2018), no studies to date have reported a robust association between TEPR and the subjective experience of effort. We speculate that a possible reason for the lack of an association between subjective and physiological measures of effort and/or tiredness from listening is that self-report measures are typically collected as a single data point "after-the-fact." Collecting data in this manner implicitly assumes that participants can accurately reflect on these subjective experiences, something which we know to be especially problematic for retrospective estimations of effort exertion (Moore & Picou, 2018; Picou & Ricketts, 2018). Further, studies in the literature have reported a potential link between TEPR and mental fatigue, particularly when examined over the course of an experimental session (Hopstaken et al., 2015; McGarrigle et al., 2017; Wang et al., 2018). It is therefore possible that the TEPR may be more closely related to changes in perceived tiredness (than with perceived effort) during an effortful listening task.

Based on our summary of the literature, we propose two potential competing accounts of the relationship between TEPR and subjective effort and tiredness from listening. The "traditional" hypothesis refers to the assumption that subjective tiredness from listening is a consequence of the repeated or sustained application of effortful cognitive processing (e.g., van der Linden et al., 2003), and TEPR can be thought of as a physiological manifestation of this effort (McGarrigle et al., 2014; Pichora-Fuller et al., 2016). In other words, if TEPR reflects transient listening effort, and if demands on capacity increase with the onset of fatigue (Hockey, 2013), then, as tiredness from listening (and/or effort) ratings increase, so too should TEPR. Alternatively, a competing hypothesis can be derived from the possibility that reduced TEPRs over time are a physiological manifestation of depleted task-related cognitive resources (Hopstaken et al., 2015; Kuchinsky et al., 2014; Wang et al., 2018), which coincides with a more pronounced

subjective experience of tiredness from listening. We refer to this as the "resource depletion" hypothesis. From this perspective, as tiredness from listening (and/or effort) ratings increase, TEPRs should *decrease*.[2]

1.1 | Experiment 1

Before examining covariance between subjective measures and TEPR, we first wanted to ensure that we could replicate a TEPR effect that is commonly reported in the literature. Therefore, the first aim of Experiment 1 was to replicate Koelewijn et al.'s (2012) effect of SNR on TEPR during speech recognition in the presence of a competing talker. The second aim was to uncover whether analysis based on the collection of multiple data points would reveal overall differences in tiredness from listening ratings as a function of SNR. In other words, would consideration of multiple self-report administrations over the course of a listening task result in enhanced sensitivity to changes in tiredness from listening than would be expected from the traditional approach of collecting just one data point at the end of a testing condition? Finally, we aimed to examine relationships between TEPR and subjective ratings of effort and tiredness from listening. For Experiment 1, participants performed a speech recognition task in the presence of a competing talker and provided subjective ratings of effort and tiredness from listening in two different SNR conditions; "easy" and "hard." The following specific predictions were made:

1. Larger overall mean TEPRs in the hard versus the easy condition, replicating the effect of SNR on TEPR during speech recognition in the presence of a competing talker (Koelewijn et al., 2012).
2. Higher effort ratings in the hard versus the easy condition, replicating similar findings in the literature (e.g., Koelewijn et al., 2012; McGarrigle et al., 2017) and higher tiredness from listening ratings in the hard versus the easy condition, reflecting the increased sustained perceptual demands of the more challenging (hard) condition and the improved sensitivity afforded by collecting multiple subjective measurements.
3. Positive correlation between overall mean TEPR and both subjective effort and subjective tiredness from listening. This is based on the prediction that tiredness from listening increases as a consequence of effortful listening, which is thought to be reflected in both TEPR and subjective rating scores (Pichora-Fuller et al., 2016).

[2]We return to (and explicitly test) these theoretical predictions in Experiment 2.

1.2 | Method

Sample size, experimental design, hypotheses, outcome measures, and analysis plan for Experiment 1 were preregistered on the Open Science Framework (https://osf.io/uk32p). Raw data, stimuli, and R scripts for analysis and plots can be found at https://osf.io/cdv2r/.

1.2.1 | Participants

Twenty-eight young adults (five male) aged 18 to 30 years took part in this experiment. This sample size was based on a power analysis conducted using G*Power (Faul et al., 2009). Koelewijn et al. (2012) reported a Cohen's *d* effect size of .5 when comparing TEPRs in the presence of a single-talker masker in listening conditions similar to the current experiment. Based on the assumption that within-subject conditions are highly correlated (say $r = .70$), a sample size of 27 participants would therefore provide an estimated power of .80 to detect a difference between these conditions if one is present at the .05 alpha error probability. To ensure that an equal number of participants were included in each of our four counterbalanced item lists (see below), we rounded the sample size to 28.

All participants were native-English speakers who reported: (a) normal or corrected-to-normal visual acuity, (b) no known eye condition, and (c) 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). All participants had normal-hearing thresholds, measured as ≤20 dB at 0.5, 1, 2, and 4 kHz in each ear. 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. They 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).

1.2.2 | Equipment

PTA testing was conducted using a Kamplex Diagnostic Audiometer AD 25. During the subsequent testing, participants were positioned 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 programed 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, and recorded for later inter-rater reliability checks.

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 0.01 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). 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.

1.2.3 | Materials

Target stimuli were IEEE 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 English accent, reading the standard phonetically balanced "Rainbow Passage" (Fairbanks, 1960). Target and masker stimuli were digitally mixed using a Matlab script (Nike, 2020) to create.wav files at 20 different SNRs ranging from +4 dB to −15 dB for each of the IEEE sentences used. These mixed files were subsequently used for the adaptive screening and listening task (described below). A random 6-s portion of the masker audio file (total file duration: 74 s) was selected for target-masker mixing. For each trial, masker onset began 2 s before target onset and ended 2 s 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 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) 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 IEEE sentences were used for the adaptive screening. Each IEEE sentence was mixed with the masker stimulus to create 10 different SNRs ranging from −6 dB to −15 dB SNR, resulting in the creation of a total of 200 mixed target-masker. wav files (20 sentences × 10 SNRs). Participants heard 20 mixed target and masker sentences, which started at −6 dB and could reach a lower limit of −15 dB. If participants responded correctly, the SNR decreased by 1 dB in the subsequent trial. If participants responded incorrectly, the SNR increased by 1 dB in the subsequent trial. An incorrect response at the upper limit (i.e., −6 dB) or a correct response at the lower limit (i.e., −15 dB) resulted in no change to the SNR in the subsequent trial (i.e., it remained at −6 dB or −15 dB, respectively). Each participant's 50% performance threshold was calculated as the mean SNR across sentences 10–20 (rounded to the nearest whole number). In cases where a "0.5" decimal value was calculated, we rounded down (e.g., −12.5 dB SNR was rounded down to −13 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 would lead to withdrawal from the task (Borghini & Hazan, 2018). Overall, the mean adapted SNR value for hard condition was −9.5 dB (SD = 1.75).

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 in the listening task. 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 for the listening task. A total of 120 IEEE sentences were used to create two target-masker lists (List 1 and List 2). IEEE sentences presented during the listening task differed from those presented in the adaptive screening. For List 1, the first 60 IEEE sentences were digitally mixed with the masker stimulus to create target-masker.wav files in the 10 possible SNRs for the easy condition (from −5 dB to +4 dB). The last 60 IEEE sentences were then digitally mixed with the masker stimulus to create a total of 60 target-masker.wav files in the 10 possible SNRs for the hard condition (from −15 dB to −6 dB). For List 2, the same 120 IEEE sentences were used, but the easy and hard condition stimuli from List 1 were swapped. Thus, the target sentences that were used in the hard condition in List 1 were used in the easy condition in List 2, and vice versa. An additional four IEEE sentences were mixed with the masker stimulus to create practice trials.

Subjective ratings

During the listening task, participants were administered three self-report rating scales. First, subjective tiredness from listening was assessed as follows;

1. How tired of listening do you feel? (100-step scale from Not at all—Extremely)

The choice of wording for this scale was taken from Picou et al. (2017) and was chosen to tap tiredness arising specifically from listening demands, as opposed to other unrelated processes (e.g., relating to visual fatigue). This measure has also been shown to have high test–retest reliability (r = .84) and excellent internal consistency (α = .91) (Picou & Ricketts, 2018). Second, subjective effort was assessed as follows;

2. How hard did you have to work to understand what was said for the previous five sentences? (100-step scale from Not at all—Extremely)

Subjective effort ratings were an adapted version of the NASA task load index item assessing mental demand (Hart & Staveland, 1988), a commonly used subjective measure of effort (Dimitrijevic et al., 2019; McGarrigle et al., 2017; Pals et al., 2019; Peng & Wang, 2019; Strand et al., 2018). Finally, we assessed subjective performance evaluation as follows;

3. How would you rate your performance accuracy on the previous five sentences? (100-step scale from Poor—Good)

Subjective performance evaluation ratings were an adapted version of the performance scales used in Moore and Picou (2018). This was included in an attempt to mitigate the possibility that participants used perceived performance evaluation as a proxy of effort (Moore & Picou, 2018).

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 midpoint 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/trial.

1.2.4 | Design and procedure

On arrival, participants were seated comfortably in the sound-treated test booth and completed Pure Tone Audiometry (PTA) testing following the British Society of Audiology recommended procedure (2011). After the PTA test, eye tracker setup and calibration began. Following the recommendations of Winn et al. (2018), soft room lighting was used and the computer screen had a grey background with reduced brightness settings (screen brightness measured at 100 cd/m^2) to minimize any visual discomfort. The seat height and/or chinrest could be 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: *"You will now perform a brief listening 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 an SNR of −6 dB. 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. After hearing and scoring the participant's response, 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 after the experimenter had 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 IEEE 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, and planks). Even minor deviations from a single key word, including inflections or derivations (e.g., "plank" instead of "planks"), were deemed to be an incorrect response. The adaptive screening lasted approximately 5 min.

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

In between blocks, participants were given the opportunity to rest inside the booth. In general, participants tended to resume the experiment within one minute. A 3-s intertrial 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 s), there was at least 4–5 s between the participant's verbal response and the recording of the subsequent trial baseline. This is consistent with Winn et al. (2018) recommended ITI of 4–6 s for experiments involving verbal responses. Each condition block lasted approximately 18–20 min. Including PTA testing, eye tracker setup and calibration, instruction period, adaptive screening, and the listening task, the total session lasted approximately 1 hr.

1.2.5 | Analysis

Pupillometry

Following the recommendations of Winn et al. (2018), pupil data were preprocessed 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., caused by 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 using R version 4.0.0 (RStudio Team, 2019) for preprocessing and analysis.

Any missing values in the data file (e.g., caused by blinks) were coded as "NA" and linearly interpolated across using values from previous and subsequent data points. Trials that contained >25% missing data were removed from the analysis. This resulted in the removal of 46 trials across all participants (1.4% of all trials in the data set). Baseline-correction was performed on each trial. The 2 s of masker speech preceding the onset of the target speech was used as the baseline window. The mean pupil size value recorded during this 2-s window was then subtracted from every sample recorded after target speech onset to provide a TEPR value. Consistent with the literature (Winn et al., 2018), we found that TEPR started to emerge approximately 1 s after target onset and peaked approximately 1 s after target offset (see Figure 2). As a result, TEPR was calculated as the relative change from baseline during the 3-s window following target speech onset. This helped to rule out any pupil size changes elicited by behavioral and/or preparatory motor responses.

A repeated-measures ANOVA was conducted to examine mean differences in TEPR as a function of condition (easy, hard) and block (1, 2, 3, 4, 5, 6). For the "rmcorr" analysis (described in more detail below), the preprocessed time series data were averaged providing a mean TEPR for every 10 trials of the 60 in each condition. By-block mean TEPR values were calculated to assess changes in TEPR over time.

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. A second independent rater transcribed the responses offline using audio recordings of each trial.[3] All discrepancies between the independent rater scores were subsequently resolved upon discussion. A repeated-measures ANOVA was conducted to examine differences in mean speech recognition performance as a function of condition (easy, hard) and the linear trend over time using block (1, 2, 3, 4, 5, 6) as a continuous factor. For the rmcorr analysis, mean speech recognition performance percentage scores were calculated every 10 trials to assess changes in performance accuracy over time.

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 that was recorded at the beginning of each block. Repeated-measures ANOVAs were conducted to examine differences in effort, performance evaluation, and tiredness from listening ratings as a function of condition (easy, hard) and the linear trend over time using block (1, 2, 3, 4, 5, 6) as a continuous factor. For the rmcorr analysis, by-block mean 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.

Correlations between measures

Correlations between TEPR, performance evaluation ratings, and tiredness from listening ratings were examined using standard Pearson's correlation tests. These tests were performed on both the overall data (i.e., collapsed across condition) and within each individual condition. The standard correlation test approach (described above) can be useful in

[3]Due to a programing error, no audio was recorded for the final trial of the first block (i.e., trial 60) for every subject in Experiment 1. As a result, scores on this particular trial could not be verified by a second independent reviewer.

determining whether there are associations between measures in terms of the overall scores that they produce. However, before conducting these tests, scores must be averaged (e.g., across conditions or time points) in order to meet the assumption of independence of error between observations; for example, there is likely to be nonindependence when sampling data from the same participants across multiple time points (Bakdash & Marusich, 2017). Aggregation of scores in this manner can disguise potentially informative intraindividual associations between these measures. An alternative approach to analyzing within-subject associations between variables that harnesses the high degree of statistical power inherent in a fully repeated-measures design is repeated-measures correlation ("rmcorr") (Bakdash & Marusich, 2017).[4] Rmcorr analysis estimates the common regression slope (i.e., the linear association shared among individuals) for two paired repeated measures, and can therefore, be a powerful statistical tool for assessing the extent to which two measures provide convergent information. All rmcorr plots and analyses were conducted in R Studio.

1.3 | Results

1.3.1 | Speech recognition performance

Figure 1 (left panel) shows speech recognition performance as a function of condition and block. There was a significant main effect of condition on performance accuracy ($F_{(1,27)} = 213.95$, $p < .001$, partial $\eta^2 = 0.89$). Overall, performance accuracy was higher in the easy ($M = 93.20\%$, $SE = 0.59\%$) than the hard ($M = 67.64\%$, $SE = 1.93\%$) condition. There was also a significant main effect of block on the linear term, ($F_{(1,27)} = 29.60$, $p < .001$, partial $\eta^2 = 0.52$), with mean speech recognition performance showing a general improvement over time. No significant difference was found between conditions in terms of the linear change over time ($F_{(1,27)} = 2.64$, $p = .12$, partial $\eta^2 = 0.09$).

1.3.2 | TEPR

Figure 1 (right panel) shows mean TEPR as a function of condition and block. Figure 2 shows the mean TEPR time series at the level of the individual trial (i.e., sentence recognition) in each condition. There was a significant main effect of condition on mean TEPR ($F_{(1,27)} = 30.51$, $p < .001$, partial $\eta^2 = 0.53$). Overall, Mean TEPR was higher in the hard ($M = 154.94$, $SE = 25.79$) than the easy ($M = 77.06$, $SE = 19.98$) condition. There was also a significant main effect of block on the linear term, ($F_{(1,27)} = 39.82$, $p < .001$, partial $\eta^2 = 0.60$), with mean TEPR showing a general decrease over time. No significant difference was found between conditions in terms of the linear change over time ($F_{(1,27)} = 0.36$, $p = .55$, partial $\eta^2 = 0.01$).

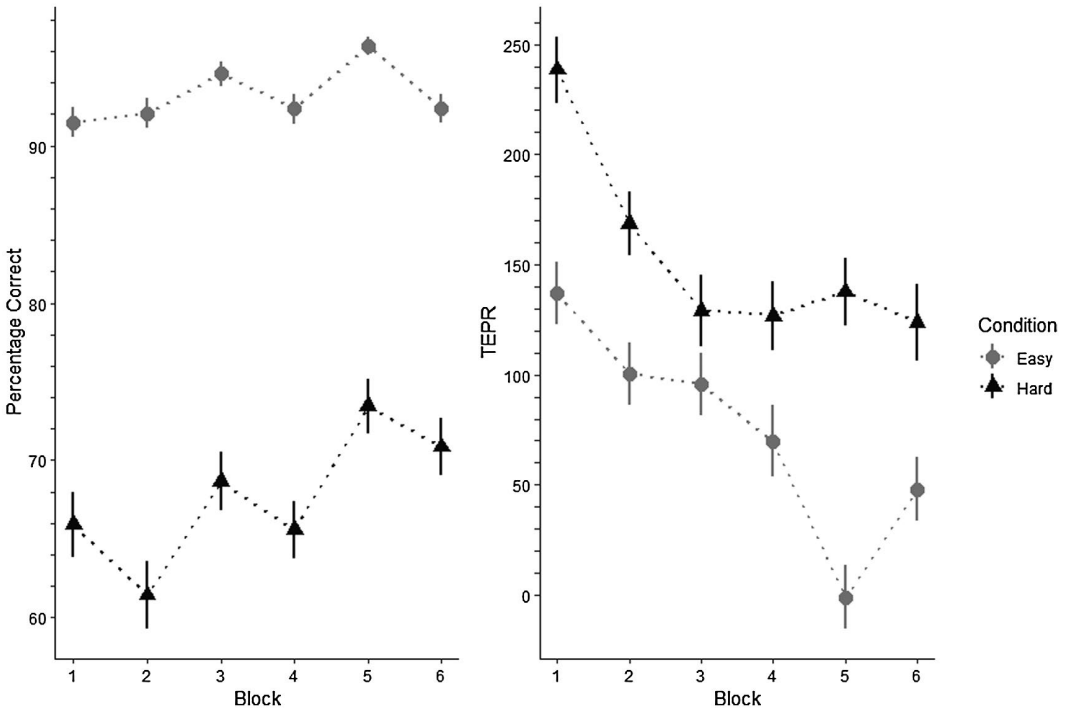

FIGURE 1 Left panel: Mean % correct speech recognition performance for each condition and block. Right panel: Mean TEPR for each condition and block. Error bars represent the standard error of the mean, *SE*

data collection and so results therein are treated as exploratory.

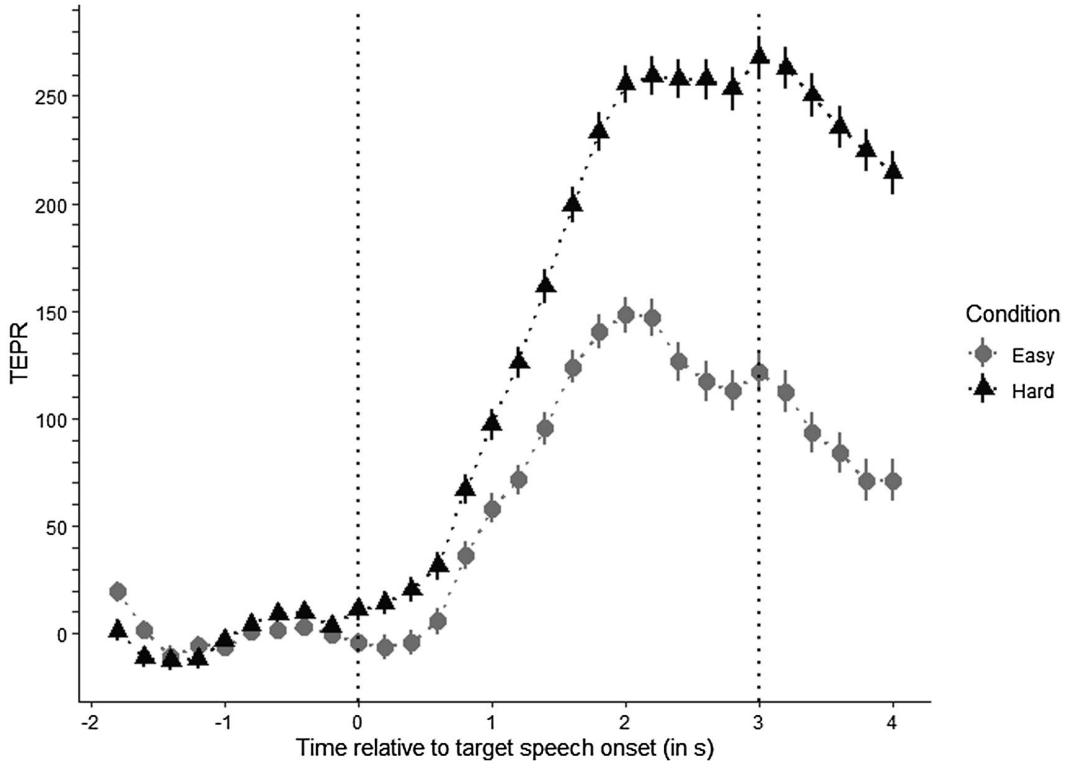

FIGURE 2 Mean baseline-corrected task-evoked pupil response (TEPR; in arbitrary units representing number of thresholded pixels) in the easy and hard conditions. Error bars represent the standard error of the mean, *SE*. Vertical dotted lines represent the beginning and end of the TEPR interval

1.3.3 | Subjective ratings

Figure 3 displays each of the three subjective rating scores (effort, tiredness from listening, and performance evaluation) in each condition as a function of block. There was a significant main effect of condition on effort ratings ($F_{(1,27)}$ = 196.78, $p < .001$, partial $\eta^2 = 0.88$). Overall, effort ratings were higher in the hard ($M = 66.49$, $SE = 2.23$) than the easy ($M = 32.28$, $SE = 2.49$) condition. There was no significant main effect of block on the linear term, ($F_{(1,27)} = 1.22$, p = .28, partial $\eta^2 = 0.04$). No significant difference was found between conditions in terms of the linear change over time ($F_{(1,27)} = 0.93$, $p = .34$, partial $\eta^2 = 0.03$).

There was a significant main effect of condition on tiredness from listening ratings ($F_{(1,27)} = 5.35$, $p = .03$, partial η^2 = 0.17). Overall, tiredness from listening ratings were higher in the hard ($M = 16.85$, $SE = 2.69$) than the easy ($M = 7.07$, $SE = 2.77$) condition. There was a significant main effect of block on the linear term, ($F_{(1,27)} = 81.09$, $p < .001$, partial η^2 = 0.75), with mean tiredness from listening ratings showing a general increase over time. There was also a significant difference between conditions in terms of the linear change over time ($F_{(1,27)} = 4.34$, $p = .05$, partial $\eta^2 = 0.14$). Tiredness from listening ratings showed a more steeply rising increase over time in the easy condition than in the hard condition.

Finally, there was a significant main effect of condition on performance evaluation ratings ($F_{(1,27)} = 147.66$, $p < .001$, partial $\eta^2 = 0.85$). Overall, performance ratings were higher in the easy ($M = 68.13$, $SE = 2.40$) than the hard ($M = 35.95$, $SE = 2.50$) condition. There was no significant main effect of block on the linear term, ($F_{(1,27)} = 3.72$, $p = .06$, partial η^2 = 0.12). No significant difference was found between conditions in terms of the linear change over time ($F_{(1,27)} < 0.001$, $p = .99$, partial $\eta^2 < 0.001$).

1.3.4 | Correlations

Standard

Standard Pearson's *r* (or Spearman's rho) correlation tests were conducted to examine relationships between each of the different measures both overall (i.e., collapsed across condition) and within each condition. These analyses were conducted on data collapsed across blocks. A total of 12 correlation tests were conducted to test our hypotheses, resulting in a Bonferroni-corrected alpha criterion significance level of .004 (.05/12). We found no significant correlation between overall mean effort ratings and overall mean tiredness from listening ratings, r_s = .38, $p = .05$, as well as no significant correlation between effort ratings and tiredness from listening ratings within the easy

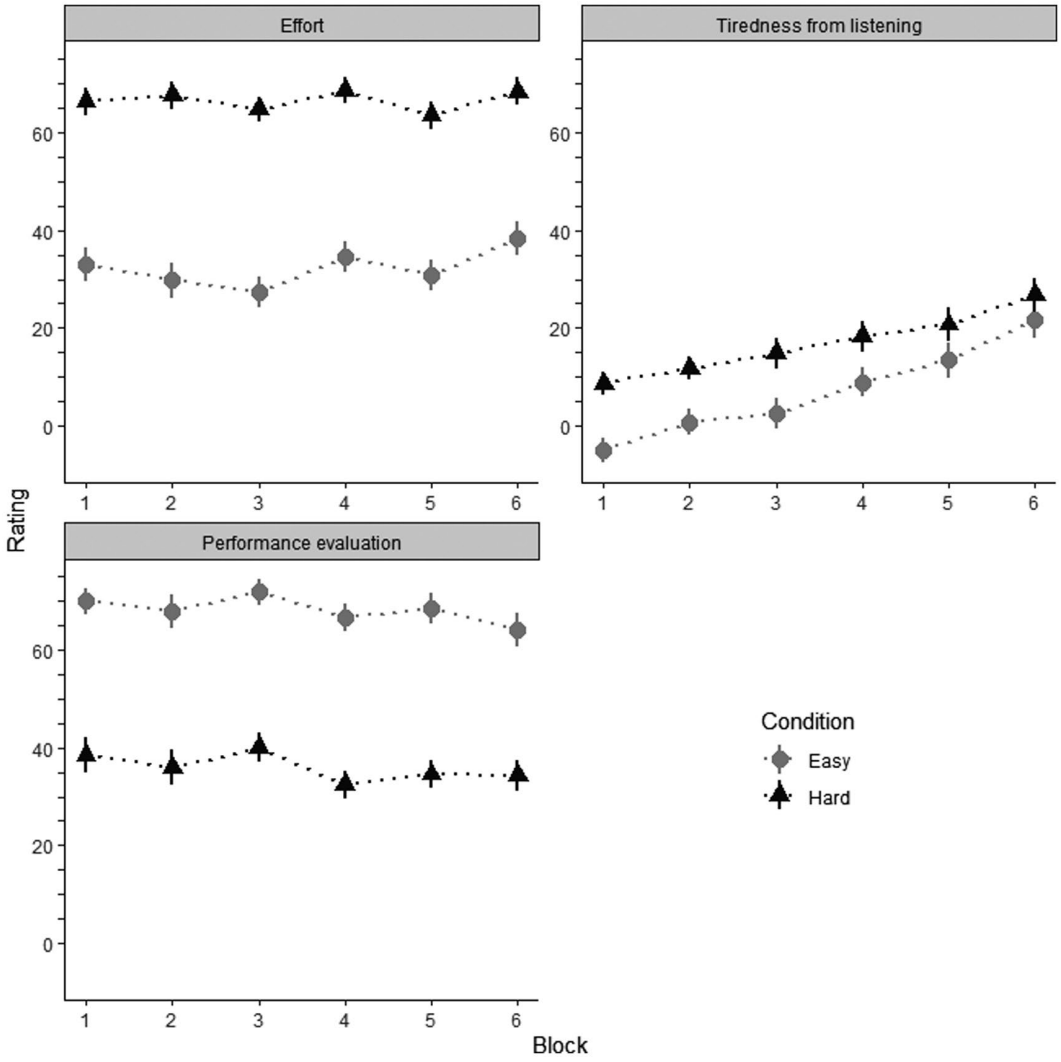

FIGURE 3 Mean subjective ratings (0–100 scale) for the easy and hard conditions as a function of block. Tiredness from listening ratings were calculated as the relative change from a baseline recorded at the beginning of block 1. Error bars represent the standard error of the mean, *SE*

condition only ($r_s = .05$, $p = .82$) or the hard condition only ($r_s = .27$, $p = .16$). Mean TEPR did not correlate with effort ratings overall ($r_s = -.17$, $p = .40$) or within each condition (easy; $r_s = .02$, $p = .90$, hard; $r_s = -.16$, $p = .42$). Mean TEPR also did not correlate with tiredness from listening ratings overall ($r = -.25$, $p = .21$) or within each condition (easy; $r = -.32$, $p = .10$, hard; $r = .05$, $p = .81$). Finally, mean TEPR did not correlate with speech recognition performance overall ($r_s = .08$, $p = .67$) or within each condition (easy; $r_s = -.28$, $p = .15$, hard; $r_s = -.13$, $p = .51$).

Rmcorr

Rmcorr analyses were conducted to explore associations between each of the dependent variables at the intraindividual level. Six "block" values were therefore collected for each participant and each dependent variable to represent change over time. As with the standard correlation tests, we examined relationships both overall (i.e., collapsed across conditions) and within each condition. We examined all possible relationships between each of the five dependent variables (effort ratings, tiredness from listening ratings, performance evaluation rating, speech recognition performance, and TEPR), resulting in a total of 30 correlation tests. We therefore applied a Bonferroni-corrected alpha criterion significance level of .001 (.05/30).

Table 1 shows rmcorr coefficients for within-subject correlation tests between all outcome measures. Rmcorr yielded a positive relationship between overall mean effort ratings and overall mean tiredness from listening ratings. Higher effort ratings were associated with higher tiredness from listening ratings. Condition-specific analyses revealed that this association was significant in the easy, but not the hard, condition. Changes in overall mean TEPR showed a negative correlation with changes in overall mean tiredness from listening ratings. Smaller TEPRs coincided with increased tiredness from listening ratings. Condition-specific analyses revealed that this association was significant in the easy, but not the hard, condition. However, changes

TABLE 1 Rmcorr correlation coefficients (and 95% confidence intervals) for within-subject correlation tests between all outcome measures

	1	2	3	4
Overall				
1. TEPR				
2. Effort rating	.03 [−.14, .19]			
3. Tiredness from listening rating	**−.40** [−.53, −.25]	**.37** [.21, .50]		
4. Performance evaluation rating	.09 [−.07, .26]	**−.70** [−.78, −.61]	**−.42** [−.55, −.27]	
5. Speech recognition performance	−.25 [−.40, −.08]	**−.31** [−.45, −.15]	.19 [.03, .35]	.24 [.08, .39]
Easy condition only				
1. TEPR				
2. Effort rating	.06 [−.11, .23]			
3. Tiredness from listening rating	**−.29** [−.43, −.13]	**.39** [.24, .52]		
4. Performance evaluation rating	.05 [−.22, .12]	**−.76** [−.83, −.69]	**−.36** [−.49, −.20]	
5. Speech recognition performance	−.19 [−.35, −.03]	**−.42** [−.55, −.28]	.01 [−.16, .18]	**.48** [.34, .60]
Hard condition only				
1. TEPR				
2. Effort rating	.11 [−.06, .27]			
3. Tiredness from listening rating	−.20 [−.36, −.04]	.19 [.03, .35]		
4. Performance evaluation rating	.04 [−.20, .13]	**−.68** [−.76, −.58]	**−.28** [−.42, −.11]	
5. Speech recognition performance	−.15 [−.31, .02]	**−.42** [−.55, −.27]	.07 [−.10, .23]	**.41** [.26, .54]

Note: Coefficients in bold are significant at the Bonferroni-corrected alpha criterion of $p < .001$.

in mean TEPR did not correlate with changes in mean effort ratings overall, nor within each condition. Significant negative associations were found between mean tiredness from listening ratings and mean performance evaluation ratings overall and within each condition. Tiredness from listening ratings generally increased as performance evaluation ratings decreased. Speech recognition performance showed a significant negative association with effort ratings both overall and within each condition. Performance improvements were generally associated with reductions in perceived effort ratings. And finally, significant negative associations were found between effort and performance evaluation ratings both overall and within each condition. Effort ratings generally increased as performance evaluation ratings decreased. All other correlation test results were nonsignificant ($ps > .001$).

1.4 | Discussion

For Experiment 1, the primary objectives were to: (a) replicate the effect of SNR on TEPR during a competing talker task (Koelewijn et al., 2012), (b) examine whether subjective effort and tiredness from listening ratings also change as a function of SNR, and (c) test for associations between TEPR and subjective ratings of effort and tiredness from listening. First, participants showed a larger TEPR in the hard than the easy condition, replicating Koelewijn et al. (2012). This suggests that a 10 dB reduction in SNR elicits an increase in the allocation of cognitive resources required to understand speech in the presence of a competing talker. This primarily served as a manipulation check and helped to ensure that we were examining a well-established pupillometry effect. Second, we found an effect of SNR on both subjective effort and tiredness from listening ratings, with higher ratings recorded in the hard versus the easy condition across both measures. Higher overall effort ratings in the hard versus the easy corroborates findings in the literature, clearly demonstrating an effect of SNR on subjective effort ratings (McGarrigle et al., 2017; McMahon et al., 2016; Rennies et al., 2014; Seeman & Sims, 2015; Strand et al., 2018; Zekveld et al., 2010). Higher overall tiredness from listening ratings in the more adverse (i.e., negative) SNR condition supports Picou et al. (2017), but not McGarrigle et al. (2017). This discrepancy may relate to the methodology used; both the present study and Picou et al. (2017) used a scale that specifically assessed tiredness from listening, whereas McGarrigle et al. (2017) administered the domain-general Visual Analog Scale for Fatigue (VAS-F) to examine differences in listening-related fatigue. It is possible that the tiredness from listening scale is more sensitive to the kinds of challenges posed by adverse SNRs. However, it is also noteworthy that the effect of SNR emerged only when data were aggregated across an entire block. In other words, the traditional approach of administering a questionnaire pre and post manipulation would

have likely revealed no such effect of SNR, as only the final data point would have been entered into the analysis (cf. Figure 3). This suggests that perceived tiredness/fatigue may show differences in fluctuation patterns as a function of SNR, and highlights the importance of administering self-report scales on a continuous basis to capture potentially subtle differences in perceived tiredness from listening.

The difference in tiredness from listening ratings between the easy and hard conditions appeared to reduce over time (see Figure 3). This pattern of change is somewhat unexpected; although no study to our knowledge has specifically investigated this phenomenon, it would be intuitive to predict that tiredness from listening might show a steeper linear increase over time, reflecting the heightened demands of sustained effort, in more-challenging listening conditions (cf. Hornsby, 2013; McGarrigle et al., 2017). One possible interpretation for the observed data could stem from changes over time in the relative contributions of perceived duration and task demand. In other words, perceived demand (i.e., how adverse the SNR is) and duration (i.e., how long the task feels) both likely influence our own subjective tiredness judgments. However, the relative contribution of each may change as a function of time such that duration becomes more salient as the task progresses, thus, mitigating the relative influence of task demand. It should also be noted that, even in the hard condition, mean tiredness from listening ratings did not exceed 30/100 (see Figure 3). This suggests that, although the hard condition was found to be more tiring than the easy condition overall, individuals did not report particularly high levels of tiredness from listening.

Finally, correlation tests between each of the primary dependent variables yielded no significant associations. On the contrary, exploratory "rmcorr" analyses revealed significant within-subject associations between several outcome measures (see Table 1). In particular, a negative within-subject association was found between overall TEPR and tiredness from listening, but not TEPR and effort; reduced TEPRs were associated with increased tiredness from listening ratings, but no change in effort ratings. This suggests that changes over time in TEPR are more closely related with the perceptual experience of tiredness from listening than with effort. Further, a positive within-subject association was found between overall effort and tiredness from listening ratings, lending support to Hockey's (2013) model of fatigue which proposes that one's evaluation of demands on capacity (i.e., effort rating) changes dynamically with the onset of fatigue. The finding of a relationship between subjective effort and tiredness also corroborates Alhanbali et al. (2017) who reported a significant positive relationship between effort and fatigue ratings.

Rmcorr analysis also revealed a significant negative within-subject association between performance evaluation ratings and both effort and tiredness from listening ratings. Generally, effort and tiredness from listening ratings went up as performance evaluation ratings went down. The significant association between effort and performance evaluation provides further support for Moore and Picou's (2018) assertion that effort ratings at least partly reflect the more intuitive evaluation of one's own performance. The association between performance evaluation ratings and tiredness from listening hints at a potentially interesting relationship between tiredness and self-efficacy (i.e., belief in one's own ability to succeed). The possibility that tiredness from listening may have a cascading effect on one's own evaluation of communication success has potential implications for hearing rehabilitation strategies. For example, targeting a reduction in tiredness from listening during rehabilitation could become increasingly important if it is found to influence an individual's willingness to engage socially and "persevere" in an adverse communication setting (Smith et al., 2011).

Correlation results demonstrated differences between the associations revealed by standard (Pearson's *r*) correlation tests and rmcorr analyses. There are a number of potential reasons for these discrepancies. Standard correlation tests and rmcorr analyses are designed to test fundamentally different types of research question. In the case of the standard correlation test, the question is a "between-subject" one; for example, do people who report high subjective "effort" also show larger TEPRs? In contrast, the question examined with rmcorr analysis is of a "within-subject" nature; for example, when individuals show a larger increase in TEPR during a listening task, does this also coincide with a larger increase in effort ratings?[5] Another key difference between the two tests which likely impacted the results observed relates to statistical power. For standard correlation tests, within-subject data are often aggregated to meet statistical independence assumption requirements which reduces overall statistical power (Bakdash & Marusich, 2017). However, rmcorr retains and models this within-subject variance, resulting in increased power to detect an association where one exists (discussed in more detail in the "General Discussion").

Given the exploratory nature of the above rmcorr analyses, further examination was required to verify the associations reported above (Wagenmakers et al., 2018). Although tiredness from listening ratings were found to be negatively associated with TEPRs, mean change in tiredness from listening remained relatively low in Experiment 1 (≤20/100; see Figure 3), even toward the latter stages of the hard condition. Simulating a more sustained effortful listening task could induce more variability in tiredness from listening and

[5]Although rmcorr and standard correlation tests will often show a similar pattern, this may not always be the case (cf. "Simpson's paradox" discussion in Bakdash & Marusich, 2017).

effort ratings, and therefore, shed light on the associations between TEPR, subjective ratings of effort, tiredness from listening, and performance evaluation.

2 | EXPERIMENT 2

Experiment 2 had two primary aims: (a) to verify the intraindividual associations found in Experiment 1, and (b) to induce a larger degree of variability in the subjective rating scores by simulating a more sustained effortful listening task. In doing so, we were able to more directly test the predictions of the "traditional" versus the "resource depletion" accounts of the relationship between TEPR and subjective reports of tiredness from listening. Further, changes in speech recognition performance have been shown to influence subjective judgments of effort (Moore & Picou, 2018; Picou et al., 2017). Using a single fixed level of task demand (i.e., SNR), therefore, permits a closer inspection of possible associations between TEPR and subjective effort and tiredness from listening that are less likely to be influenced by changes in speech recognition performance. By administering a task that taxes both perceptual capacities (i.e., listening) and sustained attention, we were also able to test the predictions of Hockey's (2013) motivation control theory of fatigue which posits that fatigue influences the evaluation of demands on capacity. Therefore, the following predictions were made:

1. Changes in TEPR will either be: (a) *positively* related to changes in effort and tiredness from listening ratings (traditional hypothesis) or (b) *negatively* related to changes in tiredness from listening (resource depletion hypothesis).
2. Subjective effort ratings will be positively related to subjective tiredness from listening ratings, supporting Hockey's (2013) motivation control theory of fatigue prediction that fatigue influences one's own evaluation of demands on capacity.
3. Tiredness from listening ratings will be negatively related to speech recognition performance, supporting the idea that fatigue has a detrimental impact on task performance (DeLuca, 2005; Hockey, 2013).

2.1 | Method

Sample size, experimental design, hypotheses, outcome measures, and analysis plan for Experiment 2 were all preregistered on the Open Science Framework (https://osf.io/nya2g). Raw data, stimuli, and R scripts for analysis and plots can be found at https://osf.io/6mbk7/.

2.1.1 | Participants

Twenty healthy young adults (two male) aged 18 to 30 years took part in this study. Only participants who had not taken part in Experiment 1 were eligible to take part in Experiment 2. Hopstaken et al. (2015) reported a Pearson's r correlation of −.33 between TEPR and subjective fatigue in their study. Based on power estimates for detecting a medium effect size when using the repeated-measures correlation (rmcorr) technique with $k = 6$ (see Figure 4; Bakdash & Marusich, 2017), we calculated that a sample size of 20 participants should provide >80% power to detect an association between these variables if one is present at the standard .05 alpha error probability. All participants had hearing thresholds of ≤20 dB at 0.5–4 kHz in each ear. Otherwise, the same eligibility criteria and recruitment methods were used as in Experiment 1.

2.1.2 | Materials, design, and procedure

The equipment used, eye tracker setup, materials, design, and procedure were the same as those of Experiment 1, with the following exceptions. Participants performed the task in one condition (hard) only. This listening task included a total of 120 trials and lasted approximately 35–40 min. Participants performed the task continuously (i.e., without a break).[6] Two stimulus lists were created, and participants were randomly assigned to one of the two lists. List 1 consisted of the same 120 IEEE sentences used in Experiment 1. List 2 consisted of 120 IEEE sentences not used in Experiment 1. Based on pilot testing the new experiment among members of the lab, we decided to reduce screen brightness from 100 to 70 cd/m^2 to mitigate against the potential for participant discomfort. Two practice trials were administered, using the same two IEEE sentences as in Experiment 1's hard practice trials. All three subjective rating scales were administered after the second practice trial to establish baselines. The mean adapted SNR value for the main (hard) condition was −8.6 dB ($SD = 1.88$).

2.1.3 | Analysis

Minor differences in how outcome measures were administered and/or scored in Experiment 2 were as follows. Subjective ratings of effort and performance evaluation were administered every five trials, resulting in a total of 24 ratings on each scale. Mean effort and performance evaluation rating scores were therefore calculated by averaging over every four

[6]However, please note that the competing talker stimulus was not played continuously in the background. As in Experiment 1, the masker stimulus started at the beginning of each trial and ended just before the speech repetition prompt.

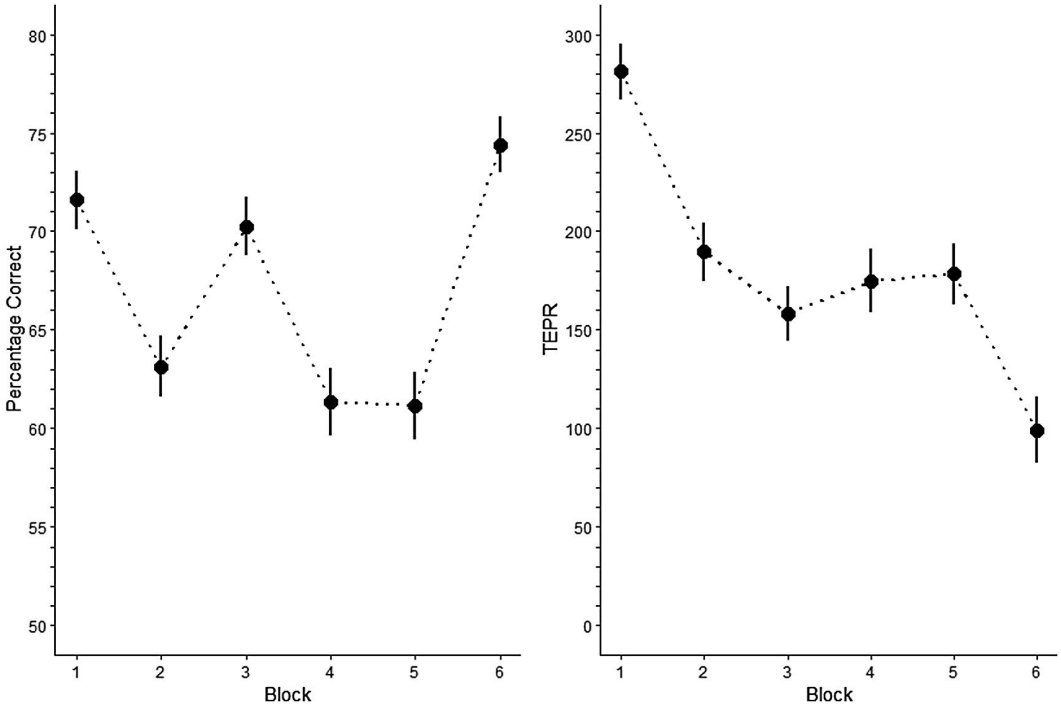

FIGURE 4 Left panel: Mean % correct speech recognition performance accuracy as a function of block. Right panel: Overall mean task-evoked pupil response (TEPR; in arbitrary units representing number of thresholded pixels) as a function of block. Error bars represent the standard error of the mean, *SE*

(rather than two) responses. For example, block 1 effort and performance evaluation ratings reflected the average effort and performance evaluation ratings as indicated after trials 5, 10, 15, and 20. Mean TEPR scores reflected TEPRs averaged over every 20 trials. Subjective ratings of tiredness from listening were administered every 20 trials (six ratings in total). A tiredness from listening subjective rating scale was administered at the very beginning of the listening task (i.e., before trial one), and this score was used as a baseline in the analysis. To summarize, each of the six blocks in Experiment 2 reflected scores averaged over 20 (rather than 10) trials. The same pupil data preprocessing techniques were used as in Experiment 1. However, on this occasion, data from one subject (s17) were removed due to having 72/120 trials with >25% missing data. Of the remaining data set, a total of 46 trials (2% of all trials in the data set) were removed from the analysis due to >25% missing sample values. One-way repeated-measures ANOVAs were conducted for each of the dependent variables to examine linear trend over time.

2.2 | Results

2.2.1 | Speech recognition performance

Figure 4 (left panel) illustrates the general pattern of change in speech recognition performance accuracy as a function of block. There was no significant main effect of block on the linear term, ($F_{(1,19)} = 0.004$, $p = .95$, partial $\eta^2 < 0.001$), with mean speech recognition performance showing no linear change over time.

2.2.2 | TEPR

Figure 4 (right panel) illustrates the general pattern of change in mean TEPR as a function of block. There was a significant main effect of block on the linear term, ($F_{(1,18)} = 35.54$, $p < .001$, partial $\eta^2 = 0.66$). Mean TEPR showed a general linear decrease over time.

2.2.3 | Subjective ratings

Figure 5 displays the general pattern of results in each of the three subjective rating scores (effort, tiredness from listening, and performance evaluation) as a function of block. For mean effort ratings, there was no significant main effect of block on the linear term, ($F_{(1,19)} = 0.65$, $p = .43$, partial $\eta^2 = 0.03$), with mean effort ratings showing no linear change over time. For mean tiredness from listening ratings, there was a significant main effect of block on the linear term, ($F_{(1,19)} = 77.61$, $p < .001$, partial $\eta^2 = 0.80$). Mean tiredness from listening ratings showed a general linear increase over time. For mean performance evaluation ratings, there was no significant main effect of block on the linear term, ($F_{(1,19)}$

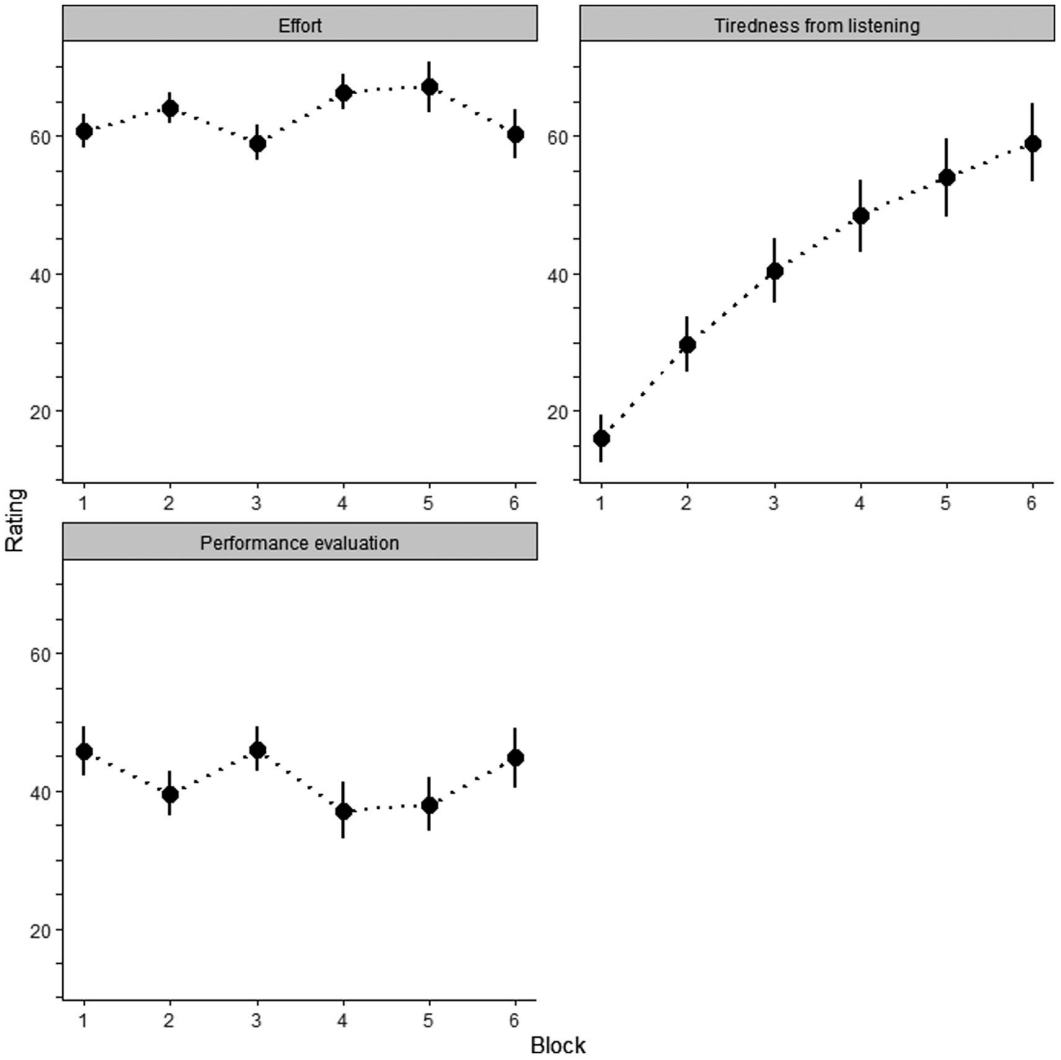

FIGURE 5 Mean subjective rating scores as a function of block. Rating scores on the *y* axis ranged from 0 to 100. Tiredness from listening rating scores were calculated as the relative change from a baseline recorded at the beginning of block 1. Error bars represent the standard error of the mean, *SE*

= 0.41, $p = .53$, partial $\eta^2 = 0.02$), with mean performance evaluation ratings showing no linear change over time.

2.2.4 | Correlations

Rmcorr

Rmcorr analyses were conducted to examine associations between the dependent variables at the intraindividual level. We examined all possible pairwise correlations between the five dependent variables (effort ratings, tiredness from listening ratings, performance evaluation rating, speech recognition performance, and TEPR), resulting in a total of 10 tests. A Bonferroni-corrected alpha criterion significance level of .005 (.05/10) was applied.

Figure 6 shows the rmcorr scatterplots pertaining to the main correlation tests of interest. Table 2 shows rmcorr coefficients for within-subject correlation tests between all outcome measures. First, changes in mean TEPR showed a significant negative correlation with changes in mean tiredness from listening ratings. Smaller TEPRs coincided with increased tiredness from listening ratings. However, changes in mean TEPR did not correlate with changes in mean effort ratings. Similarly, no significant relationship was found between changes in mean effort ratings and changes in mean tiredness from listening ratings, nor between mean TEPR and speech recognition performance. Finally, changes in tiredness from listening were not associated with either mean speech recognition performance or mean performance evaluation ratings.

2.3 | Discussion

Experiment 2 aimed to more closely examine intraindividual associations between TEPR, subjective ratings of effort and

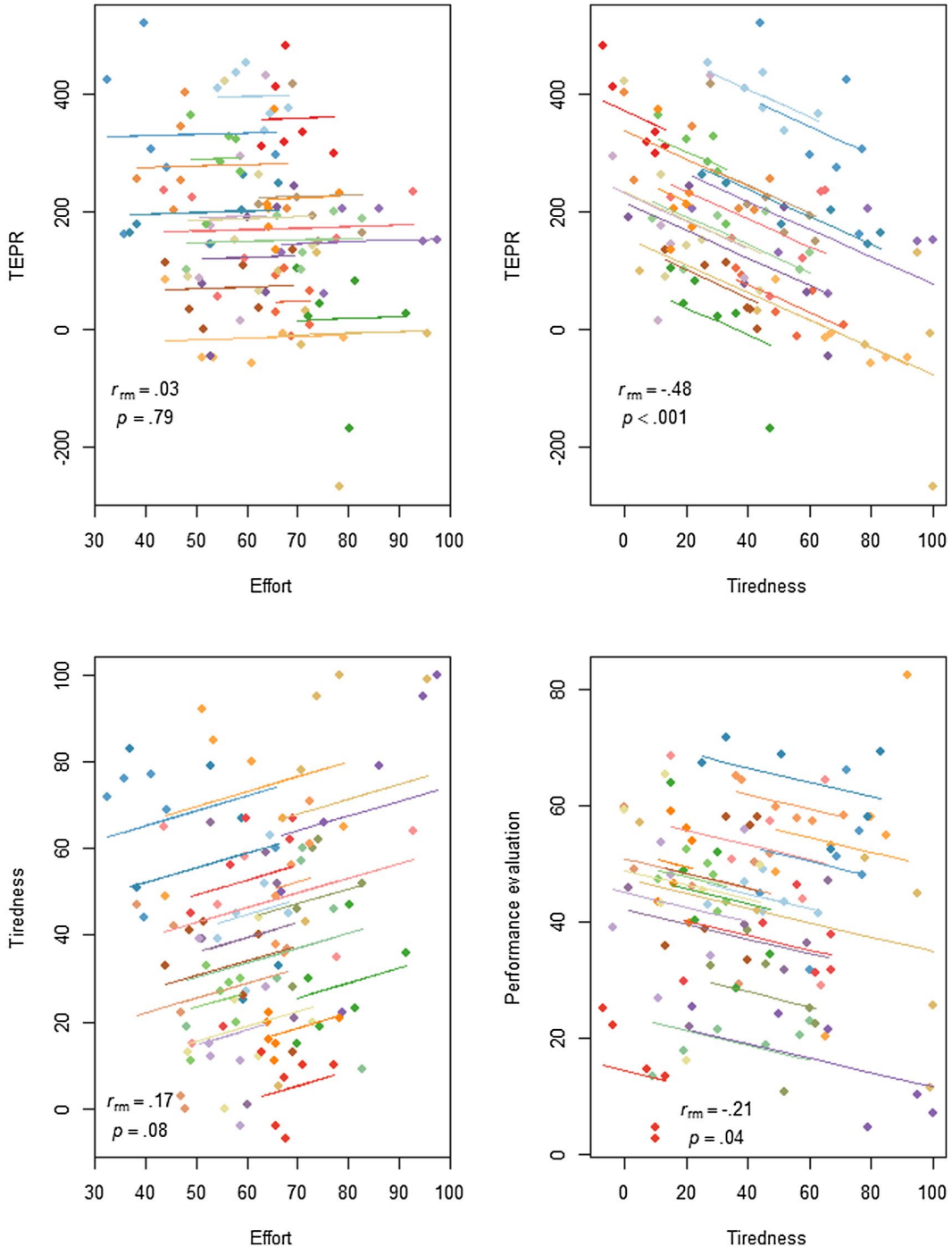

FIGURE 6 Rmcorr scatterplots showing within-subject associations between TEPR, effort ratings, tiredness from listening ratings (reported as a change from baseline), and performance evaluation ratings. Observations from a given participant are plotted in the same color, with corresponding lines showing the rmcorr fit (i.e., the common regression slope) imposed on each participant's raw data. A single data point represents the aggregate mean value for each subject on each of the six blocks. Mean data aggregation was performed over the following trials: (1) TEPR; 20 trials within each block, (2) Effort and Performance evaluation ratings; four ratings recorded after every five trials within each block, and (3) Tiredness from listening rating; a single self-report rating value at the end of each block of 20 trials

tiredness from listening, and performance evaluation. First, we found evidence in favor of the "resource depletion" account of the relationship between TEPR and tiredness from listening; TEPRs became smaller as individuals reported increased tiredness from listening. Once again, no association was found between changes in TEPR and subjective effort. Unlike Experiment 1, no significant within-subject association was found between subjective ratings of effort and tiredness from listening (possible reasons are discussed in the General Discussion). We found no significant within-subject

TABLE 2 Rmcorr correlation coefficients (and 95% confidence intervals) for within-subject correlation tests between all outcome measures

	1	2	3	4
1. TEPR				
2. Effort rating	.03 [−.18, .23]			
3. Tiredness from listening rating	**−.48** [−.63, −.31]	.17 [−.02, .36]		
4. Performance evaluation rating	.11 [−.09, .31]	**−.71** [−.80, −.60]	−.21 [−.39, −.01]	
5. Speech recognition performance	.05 [−.16, .25]	**−.49** [−.62, −.32]	.11 [−.30, .09]	**.59** [.44, .70]

Note: Coefficients in bold are significant at Bonferroni-corrected alpha criterion of $p < .005$.

association between tiredness from listening and speech recognition performance, suggesting that tiredness from listening did not have a detrimental impact on task performance (Hockey, 2013). Finally, evidence for an association between tiredness from listening and performance evaluation ratings was weaker (and nonsignificant) in this experiment ($r = -.21$) compared with Experiment 1 ($r = -.42$). Potential reasons for these discrepant results are also discussed in the General Discussion.

3 | GENERAL DISCUSSION

3.1 | TEPR as a marker of tiredness from listening, not effort

In both experiments, a negative within-subject association was found between TEPR and tiredness from listening ratings; as participants' ratings of tiredness from listening increased, their TEPRs became smaller. This effect was even stronger in Experiment 2 ($r = -.48$) than Experiment 1 ($r = -.40$), suggesting that it is exacerbated by more sustained listening demands. On the contrary, neither experiment revealed a significant association between TEPR and subjective effort ratings. This finding lends weight to the argument that TEPR is not an objective correlate of the subjective experience of listening effort (McGarrigle et al., 2014; Pichora-Fuller et al., 2016; Strand et al., 2018). Instead, within-subject changes in TEPR appear to align more closely with the subjective experience of tiredness from listening. This supports the characterization of TEPR as a broad indicator of physiological arousal that is governed by moment-to-moment changes in locus coeruleus activity which serves to maintain goal-oriented attention (Aston-Jones & Cohen, 2005). Importantly, this arousal-mediated activity appears to exhibit signs of disruption as a challenging mental task persists and becomes fatiguing (Hopstaken et al., 2015; McGarrigle et al., 2017).

The TEPR reflects a physiological response that is time-locked to a particular stimulus or event; in this case, perception of a sentence. The current study findings suggest that the strength of TEPR declines over time, and shows an association with the perception of tiredness from listening. However, as a marker of relative change from a baseline, the evoked pupil response may be influenced by underlying changes in baseline pupil size across the duration of the experiment. In other words, it is possible that the observed reduction in TEPR over time may be driven by, or at least influenced by, more low-frequency fluctuations in arousal that are not necessarily time-locked to a stimulus or event; characterized as "tonic" (as opposed to "phasic") changes in LC-mediated pupil activity (Aston-Jones & Cohen, 2005). To explore this possibility, we conducted additional exploratory rmcorr analyses testing for any associations between baseline pupil size, TEPR, and tiredness from listening ratings (plots and rmcorr estimates are provided in Supporting Information). These analyses yielded no significant associations between changes in baseline pupil size and either TEPR or tiredness from listening ratings ($ps > .05$). This suggests that both the TEPR pattern observed and the association between TEPR and tiredness from listening ratings in the current study are not likely to have been driven by changes in baseline pupil size.

Results from the current study suggest that changes in TEPR over the course of a sustained effortful listening task correspond more closely to the subjective experience of tiredness from listening which show a pattern of change that is inversely related to TEPR; as TEPRs decrease in size, tiredness ratings increase. On the contrary, within-subject changes in perceived effort appear to be more closely driven by changes in performance evaluation, with both measures showing a more constrained pattern of change over time. Figure 7 shows a hypothetical schematic illustration of how each of these four measures reflect divergent patterns of change over time during a sustained effortful listening task, based on the results of the current study.

These findings have implications for future experiments aiming to assess the impact of a specific manipulation (e.g., listening demand or an intervention) on subjective outcome measures. Previous research has shown that effort ratings are indeed sensitive to subtle (e.g., SNR) manipulations of task demand (Krueger et al., 2017). This is perhaps unsurprising when we consider the correlation between effort ratings and speech recognition performance reported both in the literature ($r = -.43$; Picou & Ricketts, 2018) and in the current study (Experiment 1; $r = -.31$). However, if the goal of a study is to

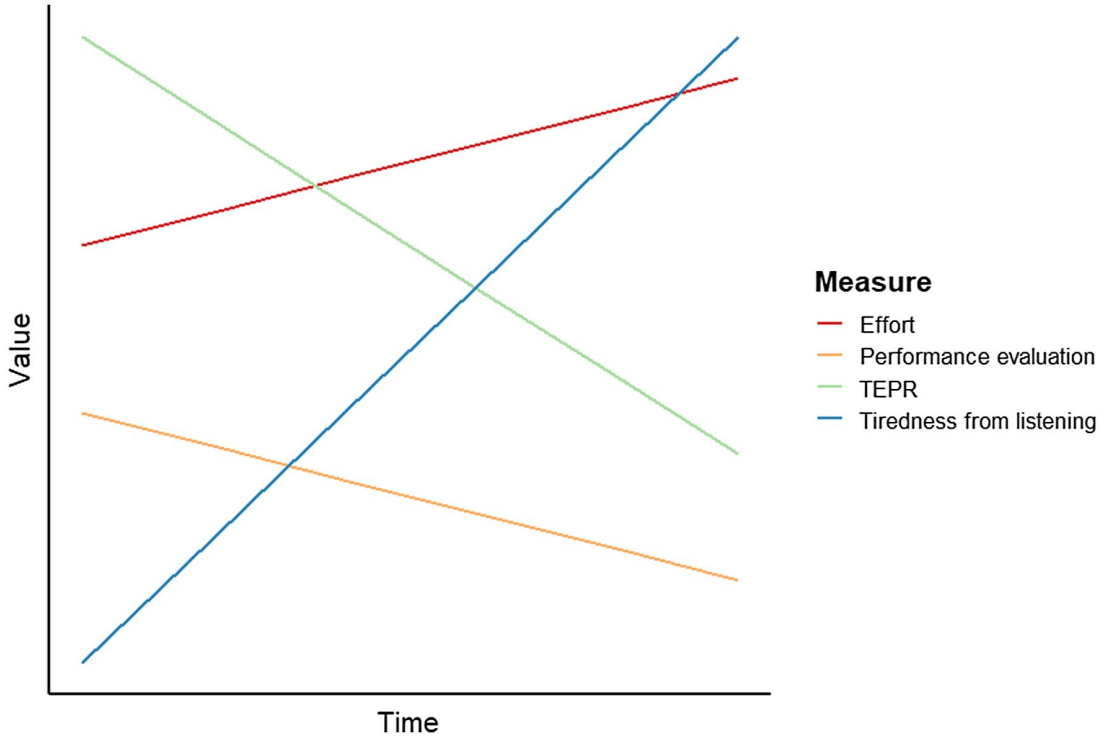

FIGURE 7 Hypothetical schematic diagram of within-subject change over time during a sustained effortful listening task for perceived effort, performance evaluation, task-evoked pupil response (TEPR), and perceived tiredness from listening. Perceived effort is relatively high (and performance evaluation relatively low) on the y axis to reflect the fixed (high) level of listening demand. Both display a similar rate of change over time, with perceived effort increases coinciding with reduced performance evaluation ratings. On the contrary, TEPR starts high (and tiredness from listening low) on the y axis to reflect early heightened levels of task engagement and arousal and the fact that fatigue has not yet started to accumulate. Both TEPR and tiredness then show a similarly steep rate of change over time (albeit in opposite directions), to reflect the more pronounced cumulative effect of time-on-task on both TEPR and tiredness from listening

examine how cognitive processes related to listening change over time, data from the current study suggest that subjective tiredness from listening ratings may be more sensitive to the effects of sustained listening demands than are effort ratings. Rather than reflecting moment-to-moment fluctuations in cognitive resource allocation, effort ratings appear to be more fixed and prone to change only in response to modulations in task demand.

It is important to emphasize that, although we found no evidence of an association between effort ratings and TEPRs, this does not mean that TEPR cannot provide useful information pertaining to effortful listening; in fact, we would argue the opposite. At the within-subject level, changes over time in physiological arousal during a listening task correspond more closely with the subjective experience of tiredness. This may (at least, partly) be because the experience of "tiredness" is simply more tractable (and therefore, easier to self-report) than the experience of "effort," which in some cases may be beyond our introspective capacities (Moore & Picou, 2018). Ultimately, this study highlights the utility of pupillometry as a measure sensitive not just to changes in task demand, but also to moment-to-moment fluctuations in tiredness from listening.

3.2 | Relations between subjective measures

It is often suggested that the mental fatigue or tiredness from listening that is reported anecdotally in hearing-impaired populations (Hétu et al., 1988; Nachtegaal et al., 2009) is likely the consequence of repeated and/or sustained episodes of effort allocation during listening (Hornsby et al., 2016; McGarrigle et al., 2014; Pichora-Fuller et al., 2016). Hockey's (2013) motivation control theory of fatigue proposed a more nuanced conceptualization of fatigue, suggesting that it serves as an emotion-like alerting mechanism which forces an individual to reevaluate their goals and priorities. In the former account, effort would be predicted to increase fatigue (Hornsby et al., 2016; McGarrigle et al., 2014; Pichora-Fuller et al., 2016). In contrast, according to Hockey's (2013) account, fatigue would be predicted to influence our effort evaluations, such that as fatigue increases, so too do effort evaluations. In either case, we would hypothesize a positive relationship between these variables. Experiment 1 revealed a significant positive association between changes in perceived effort ratings and changes in perceived tiredness from listening ratings, supporting this prediction. However, in Experiment 2, this association was no longer significant.

Closer inspection of the data from Experiment 1 revealed that the relationship (when collapsed across conditions) appears to have been driven primarily by the strength of the association in the easy condition ($r = .39$), and not the hard condition ($r = .19$). Indeed, although nonsignificant ($p = .08$), the effect size in Experiment 2 was similar to the effect size for the hard condition in Experiment 1 ($r = .17$). This suggests that the relationship between effort and tiredness from listening ratings may be contingent on the level of difficulty of the listening situation. This may result from an unexpected interaction between perceived task demand and task duration; in adverse listening scenarios, the most salient driver of effort ratings is likely the task demand (i.e., the SNR), and as long this does not change, feelings of tiredness will not likely influence effort ratings. On the contrary, in less adverse listening scenarios, a more salient factor influencing effort ratings may be the duration of the task. Increases in perceived task duration may therefore lead to a heightened sense of fatigue (Thoenes et al., 2018), thus, facilitating a stronger association between effort and tiredness from listening ratings. Future research could examine this possibility by, for example, asking individuals to provide verbal time estimates during or after the listening task.

Experiment 1 revealed a negative association between changes in tiredness from listening and changes in performance evaluation; as individuals reported increased tiredness from listening, they also tended to rate their own speech recognition performance more negatively. Although this was an exploratory finding, it hinted at the possibility of an association between an individual's subjective experience of tiredness from listening and their feelings of self-efficacy in a communication setting. While a significant association was found in Experiment 1 ($r = -.42$), a weak (and nonsignificant) association was found in Experiment 2 ($r = -.21$). Analyses of each condition subset in Experiment 1 yielded similar effect sizes for easy ($r = .36$) and hard ($r = .28$) conditions, suggesting that this effect is unlikely to be contingent on the level of task demand. Taken together, this study provides moderate evidence in favor of an association between tiredness from listening and performance evaluation. However, further research is needed to verify and explore this potential relationship.

3.3 | On the importance of (multiple) subjective ratings

In Experiment 1, we found an effect of SNR on tiredness from listening ratings, such that individuals overall reported higher tiredness from listening in the hard versus the easy condition. Visual inspection of the data (see Figure 3) suggests that this effect would have gone undetected had we used the standard approach of collecting a single data point after each condition. This highlights the importance of monitoring changes in subjective evaluations over the course of a listening task and suggests not only that the subjective experience of tiredness from listening fluctuates over the course of a listening experience, but also that this pattern of change interacts with the specific level of task demand (in this case, SNR); a steeper increase in tiredness from listening was found in the more favorable SNR in Experiment 1. However, the highest overall tiredness from listening ratings were observed in the more sustained Experiment 2 (hard) condition (see Figure 5), suggesting that tiredness from listening ratings are influenced by both task duration and perceptual demand. Changes in these kinds of subjective judgments over the course of a communication scenario will likely influence whether or not an individual chooses to withdraw or sustain engagement (Pichora-Fuller et al., 2016). Therefore, a better understanding of how these phenomena change *during*, and not just after (retrospectively), a listening task could potentially inform intervention strategies aimed at overcoming barriers to communication.

3.4 | Rmcorr

Many previous studies that have examined associations between subjective and objective measures suffer from low statistical power; a problem that is by no means confined to this particular field of inquiry (Clayson et al., 2019). Rmcorr represents a promising tool for harnessing the inherent statistical power of repeated-measures designs to examine associations at the within-subject level (Bakdash & Marusich, 2017). Studies on listening effort are typically designed to have sufficient statistical power to detect differences (e.g., in TEPR) between two or more conditions, and not necessarily associations between measures. For example, we reported an effect size of $r = -.4$ for the association between TEPR and tiredness from listening ratings. A statistical power calculation on G*Power (Faul et al., 2009) reveals that, using the standard correlation approach (i.e., Pearson's r or Spearman's rho), a minimum sample size of 46 subjects would be needed for the recommended 80% power to detect an association. Indeed, the effect size of interest in this example ($r = -.4$) is considerably larger than many effect sizes of interest in the literature, which typically fall within the small-medium range. In these cases, an even larger sample size would be required. Recruiting sufficiently large samples is not always a viable option, especially where specialist populations are concerned (e.g., cochlear implant users). As a result, powerful tests of within-subject associations like rmcorr that do not require extremely large samples can provide a practical alternative for testing within-subject associations between subjective and physiological measures.

3.5 | Study limitations and future directions

Of the 48 participants that took part in the experiments, only seven were male. We had no a priori reason to suspect that there would be sex-related differences in effortful listening given the paucity of research on this topic. However, sex differences in subjective and physiological responses are reported in the wider literature (Bath, 2020), and therefore, cannot be ruled out in the current study. To optimize generalizability, we advocate recruiting a more balanced sex ratio for future research in the area. Another limitation of the study is that we cannot infer causality from these correlational findings. For example, there are at least two potential causal interpretations for the negative within-subject association between TEPR and tiredness from listening ratings: (a) TEPR decreases as tiredness increases because fatigue is eroding the strength of the physiological response that underlies the TEPR, or (b) TEPR decreases as tiredness increases because fatigue is causing participants to disengage more readily, leading to an increasing number of non-peaking TEPR trials (and thus, a reduced mean response). Future studies could examine these potential causal mechanisms by probing for markers of task disengagement and/or distraction more explicitly. For example, this could be achieved by using paradigms that require more continuous task monitoring and engagement.

3.6 | Conclusions

This is the first study to systematically examine within-subject associations between subjective markers of effort and tiredness from listening and a commonly used physiological marker of effort (TEPR). Contrary to what is often assumed in the literature, TEPR showed a systematic within-subject association with the experience of tiredness from listening, but not effort. This study also demonstrates the importance of assessing changes in these subjective experiences over time; the effect of SNR on tiredness from listening ratings would have gone undetected using the traditional approach of collecting a single data point at the end of a listening block. Finally, we highlight the utility of assessing correlations at the within-subjects level using a highly powerful and novel analysis technique ("rmcorr"). A more detailed understanding of the subjective and physiological manifestations of effortful listening will ultimately help to mitigate this problem in various affected populations (e.g., individuals with hearing loss).

ACKNOWLEDGMENTS

We would like to thank everyone at SR research for their assistance with designing the experiments on E-Builder. We would also like to thank the technical services team (Anna, Marc, David, and Garry) in the University of York Psychology department for their help with setting up the eye-tracking lab. Thank you also to Keith Wilbraham for letting us borrow an audiometer. Thank you to Jonathan Bakdash for his valuable advice regarding the use and interpretation of "rmcorr" results. And finally, thank you to everyone in the Speech lab at the University of York (especially Sarah Knight) for useful insights and discussions. These findings were presented at: Basic Auditory Science Meeting (September, 2019), Psychonomics meeting (November, 2019), and the Speech in Noise workshop (January, 2020).

AUTHOR CONTRIBUTIONS

Ronan McGarrigle: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Supervision; Visualization; Writing-original draft; Writing-review & editing. **Lyndon Rakusen:** Investigation; Project administration. **Sven Mattys:** Funding acquisition; Resources; Writing-review & editing.

ORCID

Ronan McGarrigle https://orcid.org/0000-0003-1704-1135

REFERENCES

Agus, T. R., Akeroyd, M. A., Gatehouse, S., & Warden, D. (2009). Informational masking in young and elderly listeners for speech masked by simultaneous speech and noise. *Journal of the Acoustical Society of America*, *126*(4), 1926. https://doi.org/10.1121/1.3205403

Agus, T. R., Akeroyd, M. A., Noble, W., & Bhullar, N. (2009). An analysis of the masking of speech by competing speech using self-report data. *Journal of the Acoustical Society of America*, *125*(1), 23–26. https://doi.org/10.1121/1.3025915

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

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

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*, 456. https://doi.org/10.3389/fpsyg.2017.00456

Bath, K. G. (2020). Synthesizing views to understand sex differences in response to early life adversity. *Trends in Neurosciences*, *43*(5), 300–310. https://doi.org/10.1016/j.tins.2020.02.004

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

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

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e13437. https://doi.org/10.1111/psyp.13437

DeLuca, J. (2005). *Fatigue as a window to the brain*. MIT Press.

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

Fairbanks, G. (1960). *Voice and articulation drillbook*. Harper & Row.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

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

Gergelyfi, M., Jacob, B., Olivier, E., & Zénon, A. (2015). Dissociation between mental fatigue and motivational state during prolonged mental activity. *Frontiers in Behavioral Neuroscience*, *9*, 1–15. https://doi.org/10.3389/fnbeh.2015.00176

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). Elsevier. https://doi.org/10.1016/S0166-4115(08)62386-9

Hétu, R., Riverin, L., Lalande, N., Getty, L., & St-Cyr, C. (1988). Qualitative analysis of the handicap associated with occupational hearing loss. *British Journal of Audiology*, *22*, 251–264. https://doi.org/10.3109/03005368809076462

Hockey, R. (2013). *The psychology of fatigue: Work, effort and control*. Cambridge University Press. https://doi.org/10.1017/CBO9781139015394

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

Hopstaken, J. F., van der Linden, D., Bakker, A. B., & Kompier, M. A. J. (2015). The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics. *Biological Psychology*, *110*, 100–106. https://doi.org/10.1016/j.biopsycho.2015.06.013

Hornsby, B. W. Y. (2013). The effects of hearing aid use on listening effort and mental fatigue associated with sustained speech processing demands. *Ear and Hearing*, *34*(5), 523–534. https://doi.org/10.1097/AUD.0b013e31828003d8

Hornsby, B. W. Y., & Kipp, A. M. (2016). Subjective ratings of fatigue and vigor in adults with hearing loss are driven by perceived hearing difficulties not degree of hearing loss. *Ear and Hearing*, *37*(1), e1–e10. 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

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception and Psychophysics*, *49*(3), 227–229. https://doi.org/10.3758/BF03214307

Kidd, G., Mason, C. R., Richards, V. M., Gallun, F. J., & Durlach, N. I. (2008). Informational masking. In W. A. Yost & A. N. Popper (Eds.), *Springer handbook of auditory research: Auditory perception of sound sources* (pp. 143–190). Springer-Verlag.

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

Krueger, M., Schulte, M., Brand, T., & Holube, I. (2017). Development of an adaptive scaling method for subjective listening effort. *Journal of the Acoustical Society of America*, *141*(6), 4680–4693. https://doi.org/10.1121/1.4986938

Kuchinsky, S. E., Ahlstrom, J. B., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2014). Speech-perception training for older adults with hearing loss impacts word recognition and effort: Changes in word recognition and effort. *Psychophysiology*, *51*(10), 1046–1057. https://doi.org/10.1111/psyp.12242

Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, *1*(1), 16. https://doi.org/10.5334/joc.18

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

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

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: Physiological changes during sustained listening. *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

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

McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., Miles, K., & Graham, P. L. (2016). Monitoring alpha oscillations and pupil dilation across a performance-intensity function. *Frontiers in Psychology*, *7*, 1–12. https://doi.org/10.3389/fpsyg.2016.00745

Moore, T. M., Key, A. P., Thelen, A., & Hornsby, B. W. Y. (2017). Neural mechanisms of mental fatigue elicited by sustained auditory processing. *Neuropsychologia*, *106*, 371–382. https://doi.org/10.1016/j.neuropsychologia.2017.10.025

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

Nachtegaal, J., Kuik, D. J., Anema, J. R., Goverts, S. T., Festen, J. M., & Kramer, S. E. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. *International Journal of Audiology*, *48*, 684–691. https://doi.org/10.3758/BF03214307

Nike. (2020). *Speech in noise mixing, signal to noise ratio*. MATLAB Central File Exchange. https://www.mathworks.com/matlabcent

ral/fileexchange/37842-speech-in-noise-mixing-signal-to-noise-ratio

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

Pals, C., Sarampalis, A., van Dijk, M., & Baskent, D. (2019). Effects of additional low-pass-filtered speech on listening effort for noise-band-vocoded speech in quiet and in noise. *Ear and Hearing*, *40*(1), 3–17. https://doi.org/10.1097/AUD.0000000000000587

Peng, Z. E., & Wang, L. M. (2019). Listening effort by native and non-native listeners due to noise, reverberation, and talker foreign accent during english speech perception. *Journal of Speech Language and Hearing Research*, *62*(4), 1068–1081. https://doi.org/10.1044/2018_JSLHR-H-17-0423

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). *Ear and Hearing*, *37*, 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

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

Pisoni, D. B. (1985). Speech perception: Some new directions in research and theory. *The Journal of the Acoustical Society of America*, *78*(1 Pt 2), 381–388. https://doi.org/10.1121/1.392451

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *The Journal of the Acoustical Society of America*, *136*(5), 2642–2653. https://doi.org/10.1121/1.4897398

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*(Suppl. 2), S99–S105. https://doi.org/10.1080/14992020802301167

Rothauser, E. H., Chapman, W. D., Guttman, N., Hecker, M. H. L., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*, 225–246. https://doi.org/10.1109/TAU.1969.1162058

Rovetti, J., Goy, H., Pichora-Fuller, M. K., & Russo, F. A. (2019). Functional near-infrared spectroscopy as a measure of listening effort in older adults who use hearing aids. *Trends in Hearing*, *23*, 2331216519886722. https://doi.org/10.1177/2331216519886722

RStudio Team. (2019). *RStudio: Integrated Development for R*, Boston, MA: RStudio, PBC. http://www.rstudio.com/.

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1781–1792. https://doi.org/10.1044/2015_JSLHR-H-14-0180

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

Thoenes, S., Arnau, S., & Wascher, E. (2018). Cognitions about time affect perception, behavior, and physiology—A review on effects of external clock-speed manipulations. *Consciousness and Cognition*, *63*, 99–109. https://doi.org/10.1016/j.concog.2018.06.014

van der Linden, D., Frese, M., & Meijman, T. F. (2003). Mental fatigue and the control of cognitive processes: Effects on perseveration and planning. *Acta Psychologica*, *113*(1), 45–65. https://doi.org/10.1016/s0001-6918(02)00150-6

Wagenmakers, E.-J., Dutilh, G., & Sarafoglou, A. (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, *13*(4), 418–427. https://doi.org/10.1177/1745691618771357

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), 573–582. https://doi.org/10.1097/AUD.0000000000000512

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., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing*, *22*, 233121651877717. https://doi.org/10.1177/2331216518777174

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

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: McGarrigle R, Rakusen L, Mattys S. Effortful listening under the microscope: Examining relations between pupillometric and subjective markers of effort and tiredness from listening. *Psychophysiology*. 2020;00:e13703. https://doi.org/10.1111/psyp.13703