Investigating the Influences of Task Demand and Reward on Cardiac Pre-Ejection Period Reactivity During a Speech-in-Noise Task

Bethany Plain,[1,2] Michael Richter,[3] Adriana A. Zekveld,[1] Thomas Lunner,[2] Tanveer Bhuiyan,[4] and Sophia E. Kramer[1]

Objectives: Effort investment during listening varies as a function of task demand and motivation. Several studies have manipulated both these factors to elicit and measure changes in effort associated with listening. The cardiac pre-ejection period (PEP) is a relatively novel measure in the field of cognitive hearing science. This measure, which reflects sympathetic nervous system activity on the heart, has previously been implemented during a tone discrimination task but not during a speech-in-noise task. Therefore, the primary goal of this study was to explore the influences of signal to noise ratio (SNR) and monetary reward level on PEP reactivity during a speech-in-noise task.

Design: Thirty-two participants with normal hearing (mean age = 22.22 years, SD = 3.03) were recruited at VU University Medical Center. Participants completed a Dutch speech-in-noise test with a single-interfering-talker masking noise. Six fixed SNRs, selected to span the entire psychometric performance curve, were presented in a block-wise fashion. Participants could earn a low (€0.20) or high (€5.00) reward by obtaining a score of ≥70% of words correct in each block. The authors analyzed PEP reactivity: the change in PEP measured during the task, relative to the baseline during rest. Two separate methods of PEP analysis were used, one including data from the whole task block and the other including data obtained during presentation of the target sentences only. After each block, participants rated their effort investment, performance, tendency to give up, and the perceived difficulty of the task. They also completed the need for recovery questionnaire and the reading span test, which are indices of additional factors (fatigue and working memory capacity, respectively) that are known to influence listening effort.

Results: Average sentence perception scores ranged from 2.73 to 91.62%, revealing a significant effect of SNR. In addition, an improvement in performance was elicited by the high, compared to the low reward level. A linear relationship between SNR and PEP reactivity was demonstrated: at the lower SNRs PEP reactivity was the most negative, indicating greater effort investment compared to the higher SNRs. The target stimuli method of PEP analysis was more sensitive to this effect than the block-wise method. Contrary to expectations, no significant impact of reward on PEP reactivity was found in the present dataset. Also, there was no physiological evidence that participants were disengaged, even when performance was poor. A significant correlation between need for recovery scores and average PEP reactivity was demonstrated, indicating that a lower need for recovery was associated with less effort investment.

Conclusions: This study successfully implemented the measurement of PEP during a standard speech-in-noise test and included two distinct methods of PEP analysis. The results revealed for the first time that PEP reactivity varies linearly with task demand during a speech-in-noise task, although the effect size was small. No effect of reward on PEP was demonstrated. Finally, participants with a higher need for recovery score invested more effort, as shown by average PEP reactivity, than those with a lower need for recovery score.

Key words: cardiovascular, listening effort, monetary reward, pre-ejection period, speech reception.

INTRODUCTION

Listening can be effortful, particularly when the auditory signal is degraded, for example, by hearing loss, or in challenging acoustic environments (Plomp, 1986; Pichora-Fuller et al. 2016; Peelle, 2018). In such environments, before formulating an appropriate conversational response, several cognitively demanding processes must be undertaken. These include inhibiting any competing noise, selectively attending to the target talker, and then comprehending and memorizing semantic content of the speech (Rönnberg et al. 2008, 2013). The intentional allocation of cognitive resources to these processes is referred to as listening effort (Pichora-Fuller et al. 2016). The degree of listening effort expended by an individual is currently unmeasured and unaccounted for in clinical audiological practice, despite being a common complaint of individuals with hearing loss (Hughes et al. 2018; Holman et al. 2019). To address this, listening effort has been quantified using subjective report, as well as behavioral tasks, for example, reaction time measures (Hällgren et al. 2001) and dual-task paradigms (Wu et al. 2016). Beyond these, there is a growing interest in quantifying listening effort objectively by the use of physiological measurements (McGarrigle et al. 2014; Pichora-Fuller et al. 2016; Ohlenforst et al. 2017b).

Physiological Measures of Effort

Physiological measures that are sensitive to effort include pupil dilation, modulation of frontal and parietal alpha-band activity measured by EEG, skin conductance, heart rate variability (HRV), and cardiac pre-ejection period (PEP) (McGarrigle et al. 2014; Pichora-Fuller et al. 2016). Under controlled conditions, these measures can detect the physiological response that occurs during listening effort investment. Perhaps the most thoroughly investigated parameter illustrating effort

[1]Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology-Head and Neck Surgery, Ear and Hearing, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands; [2]Eriksholm Research Centre, Snekkersten, Denmark; [3]School of Psychology, Faculty of Health, Liverpool John Moores University, Liverpool, UK; and [4]Demant A/S, Kongebakken, Smørum, Denmark

<zdoi; 10.1097/AUD.0000000000000971>

during an auditory task is the peak pupil dilation, whereby an evoked pupil dilation is measured in response to effortful listening (Zekveld et al. 2018). The pupil size is controlled by two sets of muscles, the sphincter (constrictor) and dilator muscles, and relies on a complex interplay of parasympathetic and sympathetic nervous system innervation (Kahneman, 1973; Loewenfeld & Lowenstein, 1999; Zekveld et al. 2018). A possible downside of pupillometry is the susceptibility of the peak pupil dilation to other factors such as illumination level, as well as intrinsic factors like arousal and fatigue (Steinhauer et al. 2004; Hopstaken et al. 2015; McGarrigle et al. 2017; Wang et al. 2018a,b; Zekveld et al. 2018).

Within hearing science, heart-related physiological effort measures have been more sparsely utilized. One measure that has been successfully applied during listening is HRV, which refers to the natural variability between heartbeats (Mackersie et al. 2015; Seeman & Sims, 2015; Mackersie & Calderon-Moultrie, 2016). Notably, high-frequency HRV, which reflects parasympathetic activity, has been shown to be significantly reduced during difficult listening conditions (i.e., at difficult SNRs in individuals with hearing loss, and during faster paced speech in normally hearing individuals) (Mackersie et al. 2015; Mackersie & Calderon-Moultrie, 2016).

The focus of this study, however, was another cardiac measure that has not previously been implemented during a speech-in-noise task: the cardiac PEP. PEP differs from HRV as it is predominantly under beta-adrenergic, sympathetic nervous system control (Newlin & Levenson, 1979; Sherwood et al. 1986; Richter & Gendolla, 2009a; Richter, 2016). PEP is a systolic time interval, consisting of the period between the beginning of the electrical depolarization of the left heart ventricle and the opening of the aortic valve (Sherwood et al. 1990). This time period is related to the contractility of the heart. Sympathetic myocardial activity is coupled to effort mobilization (Wright, 1996)—a shortening of PEP is observed during increased effort investment (Richter & Gendolla, 2009b; Richter & Knappe, 2014; Krohova et al. 2017). PEP has been utilized as a measure of effort in a range of cognitively demanding tasks including mental arithmetic (Krohova et al. 2017; Mazeres et al. 2019), delayed matching-to-sample (Richter & Gendolla, 2009a), and modified Sternberg memory tasks (Richter et al. 2008). The direct link of PEP to sympathetic nervous system activity provides an advantage of this measure over other physiological measures.

Motivation and Listening Demand

Until recently, most studies investigating effortful listening have manipulated the difficulty of the listening task. Task difficulty has been manipulated by changing the signal to noise ratio (SNR), the type of interfering noise, or the pace of the talker (Mackersie & Calderon-Moultrie, 2016; Pichora-Fuller et al. 2016; Ohlenforst et al. 2017b). However, it is known that effort investment depends not only upon the difficulty of the condition, but also other contributory factors, such as a person's motivation to listen (Richter, 2016; Alhanbali et al. 2019).

Motivation is a central aspect of the Framework for Understanding Effortful Listening (FUEL) (Pichora-Fuller et al. 2016). FUEL portrays how effort investment during listening varies as a function of several factors including working memory capacity, fatigue, task demand, and motivation. When describing the interplay between listening demand and

motivation, the framework draws upon Brehm's motivational intensity theory (MIT) (Brehm & Self, 1989; Wright, 1996; Richter, 2016). MIT is centered around the principle of resource conservation, that is, minimizing the waste of energy. It states that effort investment varies in proportion to task demand, but only if success is attainable and the required effort is justified by the task's success importance. As such, MIT predicts that effort investment will directly change with the difficulty of the task until the person feels that success is impossible or no longer worth the required effort. At this point, a person will disengage from the task, thus conserving their resources and energy. In this way, success importance, or motivation, moderates the relationship between task demand and effort investment.

With FUEL and MIT in mind, several authors have investigated the moderating role of motivation on effort specifically during listening (Richter, 2016; Koelewijn et al. 2018; Zekveld et al. 2019; Zhang et al. 2019). The most common way in which these studies have attempted to manipulate motivation is by offering performance-based monetary reward during listening tasks. Offering a monetary reward is expected to increase the success importance of the task, thereby promoting engagement at higher listening demands.

The relationship between listening task demand and motivation has been investigated in several recent studies, two of which will be described here. The first, by Richter (2016), tested the relationship predicted by MIT during a tone discrimination task. During Richter's study, task demand was manipulated by presenting sine waves of either two distinct frequencies that were easy to differentiate or two closely related frequencies that were difficult to differentiate. The participants were tasked to identify whether two tones presented were identical or different. Success importance was manipulated by offering two monetary reward levels (low: CHF 0.2, or high: CHF 2) that could be obtained for 90% correctly completed trials within a block. PEP reactivity was used to show effort investment. Richter's findings were consistent with MIT, revealing that the promise of a higher reward led to greater effort investment at the difficult but not the easy condition.

The second study investigating the role of motivation on listening effort implemented a speech reception task (Koelewijn et al. 2018). Task demand was manipulated by changing the intelligibility of the sentences masked by interfering speech. Intelligibility was tracked at approximately 50 and 85% correct. Success importance was manipulated by offering a low (€0.20) or high (€5.00) monetary reward based upon 70% correct performance within a block. Listening effort was measured using the peak pupil dilation. The authors anticipated that the higher reward would have the largest impact in the more difficult condition, resulting in a greater effort investment compared to in the easy condition. Contrary to their expectations, the results showed that the higher reward level resulted in greater effort expenditure during both the easy and difficult listening conditions, but not during a control condition (speech in quiet). These two studies together provide evidence that listening-related effort investment can be moderated by changes in motivational state, as manipulated by monetary reward.

According to FUEL, two additional factors that moderate effort investment during listening are working memory capacity and fatigue (Pichora-Fuller et al. 2016). These will be introduced briefly here. A person's working memory capacity is related to their ability to store and process information (Besser

et al. 2013). Working memory capacity can be measured by tests such as the reading span test (RST) (Daneman & Carpenter, 1980; Besser et al. 2013). RST scores have been shown to relate to performance in speech-in-noise tasks. Generally, individuals with a higher RST score, which is a larger working memory capacity, perform better than those with a lower RST score during speech reception tasks with an interfering talker (Koelewijn et al. 2012; Arehart et al. 2013; Desjardins & Doherty, 2013). Previous studies have also demonstrated a relationship between working memory capacity and the magnitude of the pupil dilation response (Koelewijn et al. 2012; Wendt et al. 2016). It is not yet known if there is any impact of working memory capacity on cardiovascular responses during listening.

The influence of fatigue on cardiovascular measures during a listening task is also not yet fully established. FUEL posits that fatigue impacts subjective evaluation of performance during listening, such that higher fatigue levels decrease available resource capacity, subsequently restricting effort investment. Two recent studies have used pupillometry to investigate the relationship between fatigue and listening. For example, McGarrigle et al. (2017) revealed a reduced pupil size in the second half of each trial block compared to the first, which they interpreted as a reduction in physiological arousal. This change in arousal is experienced by the listeners as an increase in effort, representing fatigue (McGarrigle et al. 2017). Another study showed that the average peak pupil dilation during listening correlated negatively with the result of the need for recovery questionnaire (NFR), which is an index of subjective daily fatigue (Wang et al. 2018a). A higher NFR score was associated with lower effort investment overall. Yet, this has not been consistently demonstrated by others. For example, Koelewijn et al. (2018) did not find a correlation between the NFR and pupil dilation measures in a similar task performed by young adults.

Study Aims and Hypotheses

The present study investigated the effect of task demand and success importance on effort investment during a speech reception task. We aimed to build upon previous work by testing across a range of SNRs, spanning the psychometric curve from 0 to 100% correct. By doing so, we sought to obtain a more comprehensive picture of the relationship between listening effort, task demand, and success importance. Success importance was manipulated by offering two monetary reward levels. For the first time, PEP reactivity was used as an effort measure during a speech reception task. We assessed whether speech-in-noise related listening effort (as a function of task demand and moderated by success importance) is associated with changes to cardiovascular sympathetic nervous system reactivity.

In keeping with MIT and previous listening effort work (Ohlenforst et al. 2017b), it was hypothesized that effort investment would increase as the task became more difficult (i.e. at lower SNRs) until the point at which it was perceived to be impossible and participants subsequently disengaged. As PEP reactivity is a novel measure during a speech-in-noise task, it is not yet known at which SNR, or range of SNRs, disengagement will be physiologically apparent. Two conflicting hypotheses were therefore proposed and tested. On one hand, a U-shaped relationship of PEP reactivity may be demonstrated across the range of SNRs, with greater effort (more negative PEP reactivity) being seen in the medium task demands and disengagement at the most difficult conditions (Wu et al. 2016; Wendt et al.

2018). On the other hand, if the lowest SNRs presented in the present study did not yet elicit disengagement, then the relationship between PEP and SNR was expected to be linear: participants would progressively invest more effort as the task demand increased. This would be demonstrated by a more negative PEP reactivity at lower compared to higher SNRs. The relative performance of these two hypotheses will be evaluated using Bayes Factors (Masson, 2011; Mazeres et al. 2019).

The reward level was anticipated to modify the success importance of the task. At higher SNRs, it was expected that reward would not alter effort investment, as participants would perceive success to be possible and worth their energy (Brehm & Self, 1989). However, during the more difficult conditions, at the middle and lower SNRs, we predicted that the offer of a relatively high reward would increase motivation, resulting in greater effort investment compared to the low reward conditions. More specifically, for the linear hypothesis described above, we expected a more negative PEP reactivity for the high reward compared to the low reward conditions, particularly at lower SNRs. This would be demonstrated by a steeper slope when plotting SNR against PEP reactivity for the high reward conditions compared to the slope for the low reward conditions. For the quadratic hypothesis, we expected the high reward to delay disengagement compared to the low reward conditions, particularly at the lower SNRs. This would be evidenced by the U-shape (SNR against PEP reactivity) being deeper and wider, extending further to the low SNRs, for the high, compared to the low reward conditions.

An additional aim of this study was to implement and compare results from two types of PEP analysis. The first of which will be referred to here as the "block-wise" method. This method used all cardiovascular data within a block, including the participants' response window and the masking noise presentation. The block-wise method is in keeping with other studies involving PEP, that monitor changes in PEP averaged across a whole task period (Richter & Gendolla, 2009a; Richter, 2016; Mazeres et al. 2019). However, during a speech-in-noise paradigm, participants are not continuously listening to the target sentence throughout the whole block. Each trial also contains data whereby the participant was listening to the prestimulus/poststimulus masking noise, vocalizing their response, or awaiting the presentation of the next trial. It was anticipated that these additional nontarget stimuli related data may add noise to the data, or dilute any response elicited during the target stimulus listening component.

As such, the second method, which is novel to this study, involved cutting and concatenating the cycles during the presentation of the target stimulus only, when the participants were engaged in active listening. This method will be referred to as the "target stimuli" method. This method is similar to the analysis of the evoked pupil dilation response in pupillometry studies, which often exclude the prestimulus and/or poststimulus masking noise and the participant response time (Koelewijn et al. 2018; Ohlenforst et al. 2018). It was hypothesized that the target stimuli method would reduce noise in the signal, thus showing a more pronounced reactivity than the block-wise method.

The final purpose of the study was to conduct correlational analyses with the need for recovery scores and RST results. Included in the correlations were the average PEP reactivity, performance, and self-rating scores. We expected to find a correlation between a person's RST result (i.e., their working memory capacity) and their performance scores, such that participants with a better working memory capacity performed

better during the speech-in-noise task. We also aimed to determine the presence of a correlation between a participant's self-reported daily fatigue levels, as indexed by the NFR, and their average PEP reactivity. Based on the FUEL and the results of Wang et al. (2018), it was hypothesized that a greater need for recovery would be associated with a smaller PEP reactivity.

MATERIALS AND METHODS

Participants

Participants with normal hearing aged between 18 and 40 years were recruited using flyers posted at the VU University Campus and on social media. To our knowledge, this is the first study to implement PEP during a speech reception task. Because of this, there were no publications using a similar design from which to derive an estimation of effect size. Therefore, we were obliged to determine the effect size from a study using PEP during a dissimilar task (Richter & Gendolla, 2009a). A sample size of 30 was required (calculated in GPower 3.1 software), based upon an alpha error of 0.05, a power of 0.90, and an effect size of 0.22. To account for possible missing data, two additional participants were included. Therefore, a total of 32 participants (10 males, 22 females) were recruited. All participants were native Dutch speakers. The mean age of participants was 22.22 years (SD = 3.03, range = 19 to 30) and the mean body mass index (BMI) was 22.10 (SD = 2.21). Normal hearing was defined as thresholds of ≤ 20 dB HL bilaterally at 0.5, 1, 2, and 4 kHz as measured by pure tone audiometry. One participant was included with thresholds of 30 dB HL at 4 kHz only; all other frequencies were within the specified normal limits. The four-frequency pure tone average for all participants was 4.22 dB HL (SD = 4.00 dB) for the right ear and 4.73 dB HL for the left ear (SD = 4.30 dB).

The presence or absence of medical exclusion criteria was determined by self-reported medical history. Participants reported no history of psychiatric, neurological, or cardiovascular disease. More specifically, the cardiovascular diseases considered to be exclusion criteria were cardiac arrhythmia, tachycardia, severe aortic sclerosis, severe hypertension (mean arterial pressure > 130 mm Hg), aortic valve regurgitation, and defect of the septum. In addition, any history of cardiovascular procedure or intervention resulted in exclusion. Specified in the exclusion criteria were the following: the presence of an aortic prosthesis, cardiac shunt, intra-aortic pump, pacemaker, or aortic balloon. No exclusions were required based upon these medical criteria. All participants provided informed consent in accordance with the Amsterdam UMC ethical committee procedures.

Experimental Set-up

All testing was conducted in a soundproof room. Participants were seated on a chair fixed 1 m in front of a computer screen. The screen was used to present the written test instructions in Dutch. Throughout the duration of the session, the experimenter was seated approximately 1 m behind the participant at a separate computer screen displaying the cardiovascular data. The experimenter was responsible for monitoring the quality of online cardiovascular data and initiating blood pressure cuff inflation.

Procedure

Participants performed two test sessions lasting approximately one and a half hours each. They were instructed not to drink any caffeine on the day of their test sessions. Testing was conducted either in the morning or early afternoon. It was mandatory for the two sessions to be arranged on separate days, but the duration of time between the two sessions was not specified. A participant's two sessions were not required to be conducted at the same time of day. For example, some participants attended one session in the morning and one in the early afternoon, based upon when was most convenient for them each day.

During the first session, after participants signed their written informed consent, their hearing was assessed using pure tone audiometry to ensure they met the audiometric inclusion criteria. Next, in preparation for the speech-in-noise task (see details later), participants performed 12 practice sentences, consisting of two examples of each SNR that would be presented during the experiment. After this, they completed the speech-in-noise task, while being monitored using ECG and impedance cardiography (ICG) and having their blood pressure assessed, as described later.

The second session began with the speech-in-noise task. After this, each participant's near visual acuity was tested. This was to ensure that their visual abilities were sufficient to complete the RST, which is a test of working memory capacity. If the visual acuity test was not passed, the participant would be exempt from completing the RST but would otherwise remain part of the study. ECG, ICG, and blood pressure were also measured during the RST. After the RST, the participant's height and weight were measured to allow BMI to be calculated. Participants completed the NFR (see additional measures section) on paper at home in between the two test sessions. Finally, participants were carefully debriefed and the procedure for receiving their participation reimbursement was explained. All participants were reimbursed €7.50 per hour for their time plus €15.60 reward money, regardless of their performance during the task. The reward payment corresponded to the high and low reward quantities being earnt successfully in the three easiest SNRs, but not the three most difficult ones (hence, 3 × €5.00 and 3 × €0.20). Equal reward payment to all participants was a requirement stipulated by the ethical committee. Disclosure of payment rationale was not a requirement of the ethical committee, and lack of disclosure was not deemed to be problematic as participants generally performed at the expected performance levels.

Speech-in-Noise Task

Speech Stimuli and Monetary Reward • A schematic summarizing the overall structure of the speech-in-noise task is shown in Figure 1. Participants completed a speech-in-noise task with a 6 × 2 within-subject design, that is, two reward levels and six SNRs. Six fixed SNRs between –1 and –21 dB SNR, distributed in 4 dB steps, were selected as they were anticipated to span the psychometric performance curve from 0 to 100% correct for people with normal hearing, as demonstrated by Zekveld and Kramer (2014) and Ohlenforst et al. (2017). Each block contained 34 different sentences at the same SNR and lasted approximately 5 minutes in duration.

Auditory stimuli consisted of short, everyday Dutch sentences (Versfeld, Daalder, Festen, & Houtgast, 2000) delivered by a single loudspeaker placed at 0 degrees azimuth. Target speech was uttered by a female talker in the presence of a male single-talker masker consisting of concatenated sentences. Although not without limitations (see Discussion section for more detail), a single-talker masker was selected because it has previously been shown to produce the largest pupil dilation response, compared to other masking noise types (Koelewijn et al. 2012). As the sensitivity of PEP during speech-in-noise
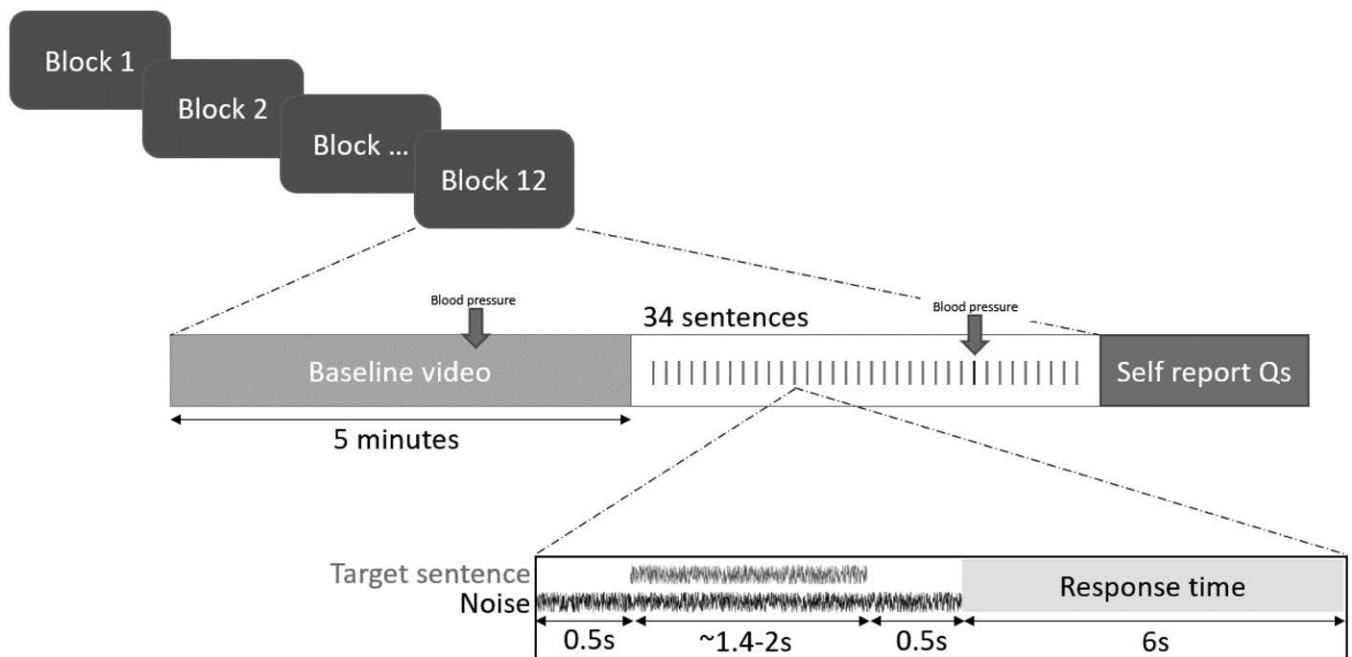
Fig. 1. Overview of the speech perception task. Qs, questions.

tasks was not known, we hoped to produce the most robust physiological response by using this masking type. The male masker speech was spectrally shaped such that the long-term average frequency spectrum was identical to that of the female target speech. The male voice preceded the female target speech by 0.5 sec and ended 0.5 sec after the conclusion of the target sentence. Target sentences lasted between 1.4 and 2.0 sec in duration (mean duration, 1.84 sec). The overall level was kept constant at 65 dB SPL; the masking noise and target levels were both adjusted to achieve this level. At the end of each sentence, participants were given 6 sec to verbally repeat the sentence spoken by the female. The following stimulus began immediately after these 6 sec. Responses were recorded using a dictaphone and were scored after the session by a native Dutch speaker. A correct score was based upon all words being correctly repeated in the sentence. Performance scores were calculated by dividing the number of correctly repeated sentences in a block by the total number of sentences. No feedback about their performance was provided to participants during the test session.

During each block of 34 sentences, participants were offered a low (€0.20) or high reward (€5.00) providing they achieved a performance score of 70% correct for that block. These reward levels were chosen based upon a previous study that also used monetary reward to motivate participants during listening (Koelewijn et al. 2018). In the present study, participants were able to see the reward level on the screen in front of them throughout the task. The 12 conditions (six SNRs at two reward levels) were divided equally between two separate test sessions. Conditions were counterbalanced* to control for order effects.

---

* The counterbalancing procedure was as follows. Participants were split equally into four groups who undertook testing in different orders. Groups A and B progressed from the easy to the most difficult SNR, whereas groups C and D progressed from the difficult to the easy SNR. The reward levels were presented to groups A and C in the following pattern (at each reward offer for three subsequent SNRs): low, high, high, and low. The opposite reward pattern was presented to groups B and D.

Baseline Measurement • A 5-minute period, that we regarded for the purpose of our analyses as 2 minutes of rest followed by 3 minutes of baseline, preceded each block of sentences. The purpose of this period was to elicit the "resting" state of the participant (Jennings et al. 1992). This allowed any change in the cardiovascular measure between conditions of the task to be determined and attributed to the task, rather than other uncontrolled factors. Throughout this whole period, participants were instructed to quietly watch a 5-minute video displayed on the screen. The videos depicted images shot through the window of a train passing through a countryside scene and contained no emotive, moving, or distressing footage. There were 12 different videos of similar content, to precede the 12 blocks of sentences across the two sessions. Participants reporting motion sensitivity were permitted to close their eyes during the video.

Self-rating Scales • Following each block of sentences, participants were required to complete self-rating scales regarding the task. The self-rating scales consisted of a printed horizontal line, numbered from 0 to 10 with markers displaying one decimal precision (i.e., 100 markers per scale). Participants circled the appropriate point on the scale for the following four items: (1) how effortful the preceding period was, (2) how difficult they found it, (3) how well they felt they performed (as an estimate of the proportion of sentences perceived correctly), and (4) their tendency to give up. The extremes of the scales for each question, respectively, were labeled as follows: (1) from "no effort" to "very effortful," (2) "not difficult at all" to "very difficult," (3) "none of the sentences were intelligible" to "all of the sentences were intelligible," and (4) "I did not give up on any of the sentences" to "I gave up on all of the sentences." These were presented to the participants in Dutch. Participants also completed self-rating scales for the first baseline video of each session. After the video, participants rated how effortful and how difficult it was to watch. The rationale for including baseline rating scales was to allow a comparison with the task-related ratings and to be able to identify any participants who may have

found the baseline video excessively effortful or difficult to watch. They were asked to answer the questions (translated to English), "How much effort did it take to watch the video?" and "How difficult did you find it [to watch the video]?". Similar to the task-related self-rating scales, the participants selected their responses on a scale from 0 to 10 with one decimal precision. The extremes of these two scales were identical to the corresponding effort and difficulty questions described above for the task-related scales.

Cardiovascular Measurements

Cardiovascular measurements were undertaken using a Cardioscreen 2000 system (Medis, Ilmenau, Germany) which measures ICG and ECG at a sample rate of 1000 Hz. These were measured using disposable solid gel ICG electrodes that were applied to the participant by the experimenter: a dual-sensor was positioned at the base of the left side of the neck, and single sensors were placed at the left middle axillary line at the level of the xiphoid process and 10 cm later.

Pre-ejection Period

Data Selection for Pre-ejection Period Analysis • The cardiovascular data recorded during the 3-minute baseline was selected for analysis. This was based upon the assumption that after the 2-minute rest period the participant should have reached a resting state. The data to be analyzed from the task period were selected using two different methods. In the first, the data were analyzed in a *block-wise* fashion. For this, PEP values were calculated from the first 3 minutes of the task period, including the target sentence presentation, the pre-stimulus and poststimulus masking noise, and the participant response window. This included around 21 trials of each block. The second method, referred to as the *target stimuli* method, involved selecting and then concatenating only the cycles during which the target sentence presentation occurred, over the whole 5-minute task period. The data from the target stimuli method excluded any data acquired when the masking noise was presented alone, when the participant was vocalizing their response, and when the participant was awaiting the next sentence presentation. The concatenated signal was around 75 sec in duration for each block (consisting of 34 target sentences, each lasting approximately 2 sec).

Deriving the Pre-ejection Period • Regardless of the method used to select the data for analysis, processing of the selected cardiovascular data to obtain the PEP were completed in the same way, using the procedure described by Richter (2016). First, R peaks in the ECG signal were detected automatically (Pan & Tompkins, 1985) and the success of automatic detection was verified by visual confirmation. Next, the ICG signal was differentiated and filtered using a low-pass Butterworth filter with a cutoff frequency of 50 Hz (order 4). Cycles containing artefacts were identified by visual inspection and subsequently excluded. Ensemble averages of 60 sec were displayed and PEP was scored in keeping with the procedure recommended (Sherwood et al. 1990): PEP was identified as the period between the R-onset of the ECG and the B-point of the ICG (see Fig. 2). For the block-wise method, each baseline and each task period generated three PEP values: one ensemble average per minute. For the target stimuli method, one PEP value per task period was generated from the concatenated data.

The next steps were to verify that PEP was being identified accurately and to calculate PEP reactivity scores. To ensure accuracy, PEP scoring was completed by two separate scorers, one of whom was blind to the experimental conditions. The scorers together reviewed any conditions for which there was a greater than 10 msec difference between their PEP scores. Any errors in scoring were corrected. The intraclass correlation coefficient (ICC) (form: two-way mixed effects, absolute agreement, and multiple raters) was computed to determine the agreement between the two PEP scorers (McGraw & Wong, 1996). For the block-wise method, the average ICC was 0.96 with 95% confidence intervals from 0.94 to 0.97 and for the target stimuli method, the ICC was 0.98 with 95% confidence intervals from 0.97 to 0.98. The average of both scorers' corrected scores was used for subsequent analysis.

The arithmetic means of the three ensemble averages (one per minute) from each baseline was calculated, such that there was a single PEP score for each baseline. For the block-wise method, the arithmetic mean of the three task-related PEP values was calculated. For the target stimuli method, there was already only one PEP value for the task period. Finally, for both methods, reactivity scores were calculated by subtracting the baseline PEP score from its corresponding task period PEP score. A shortening of PEP duration reveals increased effort investment (Richter & Gendolla, 2009b; Richter & Knappe, 2014; Krohova et al. 2017). Therefore, we expected a shorter PEP during the task than the baseline, resulting in negative reactivity scores. More negative PEP reactivity represents greater effort investment.

Heart Rate and Blood Pressure • Heart rate was determined from the inter-beat (R-R) intervals of the ECG. The Cardioscreen 2000 system also contains a blood pressure monitoring system to measure systolic blood pressure and diastolic blood pressure (DBP). The blood pressure cuff was placed over the brachial artery above the elbow on the right arm of all
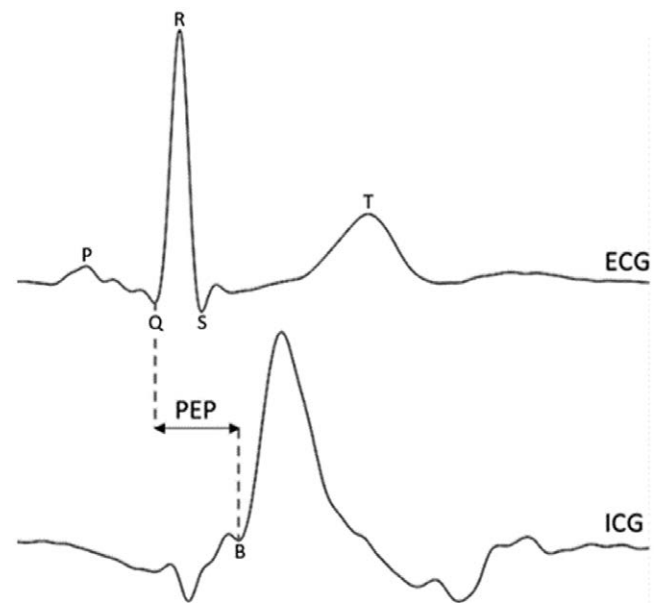

Fig. 2. Schematic demonstrating how PEP (in msec) is determined from the R onset of the ECG and the B point of the ICG. PEP, pre-ejection period, ICG, impedance cardiogram.

TABLE 1. Arithmetic means of heart rate and diastolic blood pressure reactivity

		dB SNR					
	Reward	−21	−17	−13	−9	−5	−1
HR reactivity (bpm)	High	2.52 (0.45)	1.29 (0.23)	2.72 (0.49)	2.54 (0.46)	2.66 (0.48)	3.05 (0.54)
	Low	1.55 (0.28)	1.89 (0.34)	2.39 (0.43)	2.42 (0.43)	2.69 (0.48)	3.05 (0.55)
DBP reactivity (mm Hg)	High	0.83 (0.80)	0.84 (0.90)	3.81 (0.99)	1.00 (0.97)	2.52 (0.83)	2.97 (0.95)
	Low	0.76 (0.89)	0.36 (0.84)	1.84 (1.01)	0.87 (0.67)	3.29 (0.67)	2.48 (0.95)

SEM is presented in brackets. dB SNR = decibels signal to noise ratio, bpm = beats per minute, mm Hg = millimeters of mercury.

participants, except for a single participant who had recently undergone surgery on her right arm and therefore preferred her left arm to be measured. At the onset of the 26th sentence of each block of the sentence perception test, the blood pressure cuff was inflated. It was also inflated once in the third minute of each baseline video. Reactivity scores for DBP and heart rate were calculated by subtracting the baseline values from the corresponding task values.

A prerequisite for concluding that PEP changes are due to sympathetic, beta-adrenergic activity on the heart is to examine heart rate and DBP reactivity scores, to verify that PEP reactivity scores are not influenced by preload or afterload, respectively (Obrist et al. 1987; Sherwood et al. 1990; Richter et al. 2008). If preload had contributed to negative PEP reactivity scores, this would have been demonstrated by a decrease in heart rate. If afterload had contributed to negative PEP reactivity scores, this would have been demonstrated by a decrease in DBP. As displayed in Table 1, in our dataset, negative PEP reactivity scores (for the target stimuli data) were accompanied by slight increases in both heart rate and DBP. These changes suggest that preload and afterload did not cause negative PEP reactivity scores (Obrist et al. 1987; Sherwood et al. 1990). Consequently, changes in PEP reactivity presented in the results can be justifiably interpreted as reflecting changes in cardiac sympathetic activity.

Additional Measures

Reading Span Task • The RST is a test of working memory (Daneman & Carpenter, 1980; Besser et al. 2013). During the task, participants were presented with 12 sets of sentences containing 3, 4, 5, or 6 sentences sequentially. Sentences were displayed on the screen in three parts, containing either one or two words on the screen at a time. Half of the presented sentences were semantically correct, and the other half were incorrect. Participants were instructed that after the presentation of each sentence they were to report if the sentence was semantically correct or incorrect, by stating "true" or "false," respectively. Then, at the end of each set of sentences (i.e., 3, 4, 5, or 6 sentences), participants were instructed to recall either the first or last words of each of the sentences in that set. Participants were therefore required to simultaneously assess and announce the semantic correctness of each sentence, whilst also attempting to memorize the first and last words of each sentence. The final score consisted of the total number of words correctly repeated across all the sets (a maximum of 54 words). To allow for cardiovascular reactivity scores to be assessed for the RST, another 5-minute "rest" video was displayed before the start of the RST, this time consisting of a drone-filmed aerial view of scenery. Next, the instructions for the RST were given, and participants completed three practice sentences before performing the

RST. The cardiovascular data recorded during the RST are not detailed here as these are beyond the scope of this article.

Need for Recovery Scale • The NFR was developed and validated to measure the need for recovery of working individuals (Cronbach's alpha = 0.88) (van Veldhoven & Broersen, 2003). The scale has been included previously in listening effort studies as an analog of fatigue (Koelewijn et al. 2018; Wang et al. 2018). It consists of 11 items, each of which is answered by confirmation or rejection of the item. Examples of items translated into English are: "Often after a day's work I feel so tired that I cannot get involved in other activities" and "By the end of the working day, I feel really worn out." Students were instructed to consider their need for recovery after a day at university when completing the questionnaire. For each participant, the number of items answered "yes" was determined and divided by 11 (the total number of items). This answer was multiplied by 100 to reveal a percentage NFR score for each individual.

Statistical Analysis

Two-way repeated measures analyses of variance (ANOVAs) were undertaken to determine the presence of main effects of and/ or interactions between reward (high, low) and SNR (six levels) on performance scores, self-rating scales, PEP baseline, and PEP reactivity scores. The two hypotheses regarding SNR were tested: (1) the U-shaped, or quadratic hypothesis, where disengagement was elicited at the lower SNRs and (2) the linear hypothesis, where no disengagement was elicited and progressively more effort would be invested as the task demand increased. This comparison of the two hypotheses was achieved by computing Bayes factors (Glover & Dixon, 2004; Masson, 2011; Mazeres et al. 2019).

Next, averages for each individual were taken across all conditions for self-ratings, performance, and target stimuli PEP reactivity. Two-tailed Pearson's correlation coefficients were conducted to assess the relationship between RST results and NFR scores with the average self-ratings, performance, and PEP reactivity. The Shapiro-Wilks normality test revealed that average PEP reactivity was non-normally distributed, therefore a Spearman's rank correlation coefficient was conducted for the average PEP reactivity and the NFR scores. No corrections were applied for multiple comparisons.

RESULTS

The data of one participant were excluded from all analyses due to the poor quality of the cardiovascular signal, which rendered it impossible to reliably score the B-point of the ICG. Data from the remaining 31 participants were analyzed unless otherwise stated. Where assumptions of sphericity were violated, Greenhouse-Geisser corrected p values, and degrees of freedom (df) are presented.
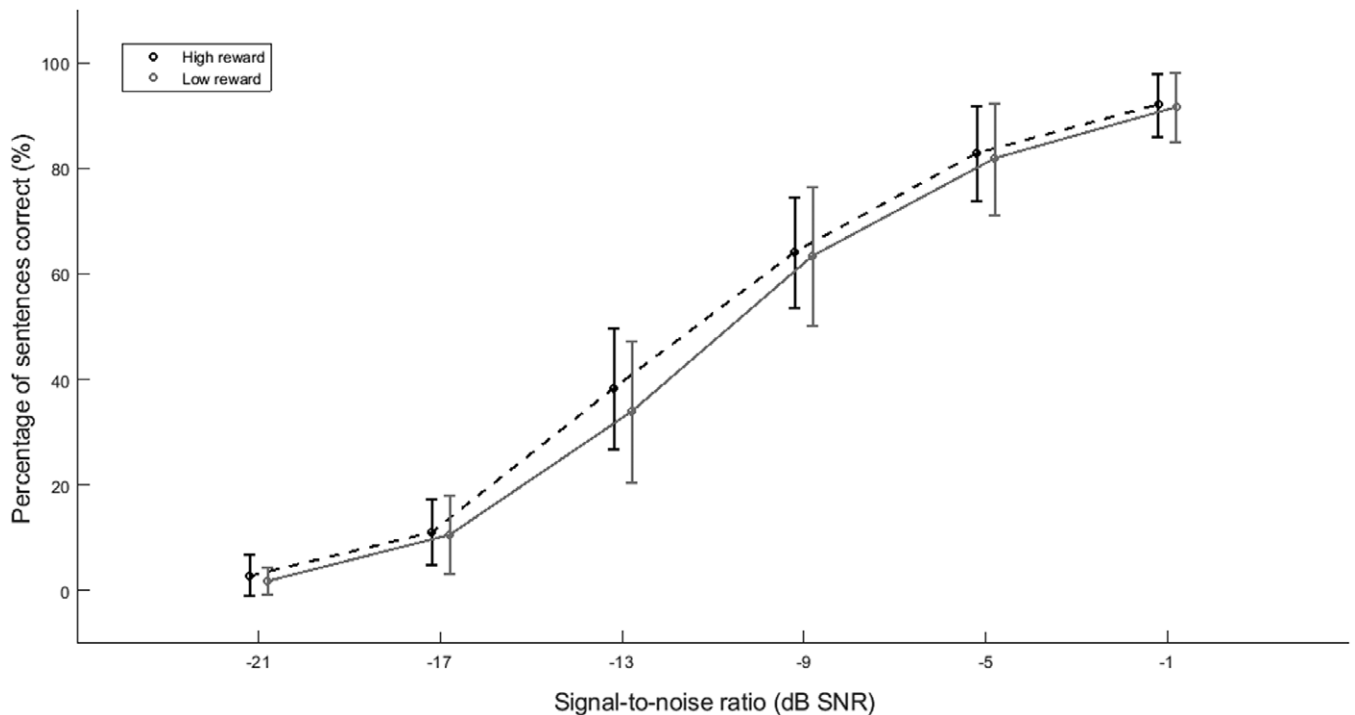
Fig. 3. Average performance scores at each signal to noise ratio, presented as a percentage of total sentences correct. Error bars represent SEM.

Performance Data

Two participants' speech-in-noise task performance data were lost due to an audio recording failure during their test sessions. Consequently, performance data from 29 of the total 31 participants were analyzed. Figure 3 shows the average performance scores based upon correct sentence recognition at each SNR. Average sentence perception scores for both high and low reward ranged from 2.73 to 91.62% from –21 dB SNR to –1 dB SNR, respectively, with lower SNRs reflected in lower scores. A two-way repeated measures ANOVA revealed a significant main effect of SNR ($F_{[2.98,83.53]} = 1297.14$, $p < 0.001$, $\eta p^2 = 0.98$), such that poorer performance was confirmed in the lower SNRs compared to the higher SNRs. In addition, a significant main effect of reward ($F_{[1,28]} = 4.63$, $p = 0.04$, $\eta p^2 = 0.14$) was demonstrated, revealing better performance in the high reward compared to the low reward conditions. No interaction between SNR and reward was found ($F_{[3.28,91.86]} = 0.77$, $p = 0.53$, $\eta p^2 = 0.03$).

Average word perception scores ranged from 16.79% (SEM = 1.26) to 98.10% (SEM = 0.35) from –21 dB SNR to –1 dB SNR. To provide an indication of whether participants were erroneously substituting words occurring in the background masker, the number of additional, nontarget words spoken by the participant were calculated for each SNR. At the lowest SNR (–21 dB SNR), on average participants spoke only 1.16 additional, nontarget words per sentence (SD = 0.99, range = 0.15 to 4.91). For comparison, at the highest SNR (–1 dB SNR) on average participants spoke even fewer: just 0.10 additional, nontarget words were spoken per sentence (SD = 0.08, range = 0 to 0.30).

Cardiovascular Data

Cardiovascular data from 31 participants were included in the analysis. As described in Methods section, PEP data were analyzed in two ways: the block-wise method and the target

stimuli method. Data from both methods are reported here. PEP reactivity scores were not correlated with BMI, which was therefore not included in the subsequent analysis.

Baseline Pre-ejection Period Data

Baseline PEP data were taken from the 3-minute baseline period. Analysis of the absolute baseline PEP data revealed no significant main effect of SNR ($F_{[1.82,54.72]} = 1.21$, $p = 0.31$, $\eta p^2 = 0.04$) or reward ($F_{[1,30]} = 0.03$, $p = 0.86$, $\eta p^2 = 0.001$), and no interaction between the two ($F_{[1.36,40.89]} = 0.05$, $p = 0.99$, $\eta p^2 = 0.002$).

Task-related Pre-ejection Period Results

Method 1: Block-wise Pre-ejection Period • The PEP data were first analyzed in a block-wise fashion. The average PEP reactivity scores and the SEMs are plotted in Figure 4. For both the high and low reward conditions PEP reactivity was negative at –21 dB SNR and became positive at the subsequent four SNRs. Finally, at –1 dB SNR, the low reward condition demonstrated a negative PEP reactivity compared to the high reward condition, which remained positive. A positive PEP reactivity reveals a shorter PEP during the baseline than the task. A repeated measures ANOVA was conducted, which demonstrated no significant main effects of SNR ($F_{[3.23,96.78]} = 2.01$, $p = 0.11$, $\eta p^2 = 0.06$), reward ($F_{[1,30]} = 0.25$, $p = 0.62$, $\eta p^2 = 0.008$) or interaction effects ($F_{[3.87,116.12]} = 1.22$, $p = 0.31$, $\eta p^2 = 0.04$). Then, one-tailed quadratic and linear within-subject contrasts were conducted to test our two alternative predictions regarding the impact of SNR on PEP reactivity. Curiously, the upside-down U-shaped pattern observed in the data was inverted with respect to our hypothesis. However, no significant effect of SNR was revealed by the quadratic contrast ($F_{[0.65,19.35]} = 5.42$, $p = 0.98$, $r = 0.15$) or the linear contrast ($F_{[0.65,19.35]} = 1.33$, $p = 0.12$, $r = 0.04$). As neither contrast demonstrated a significant effect of SNR, no
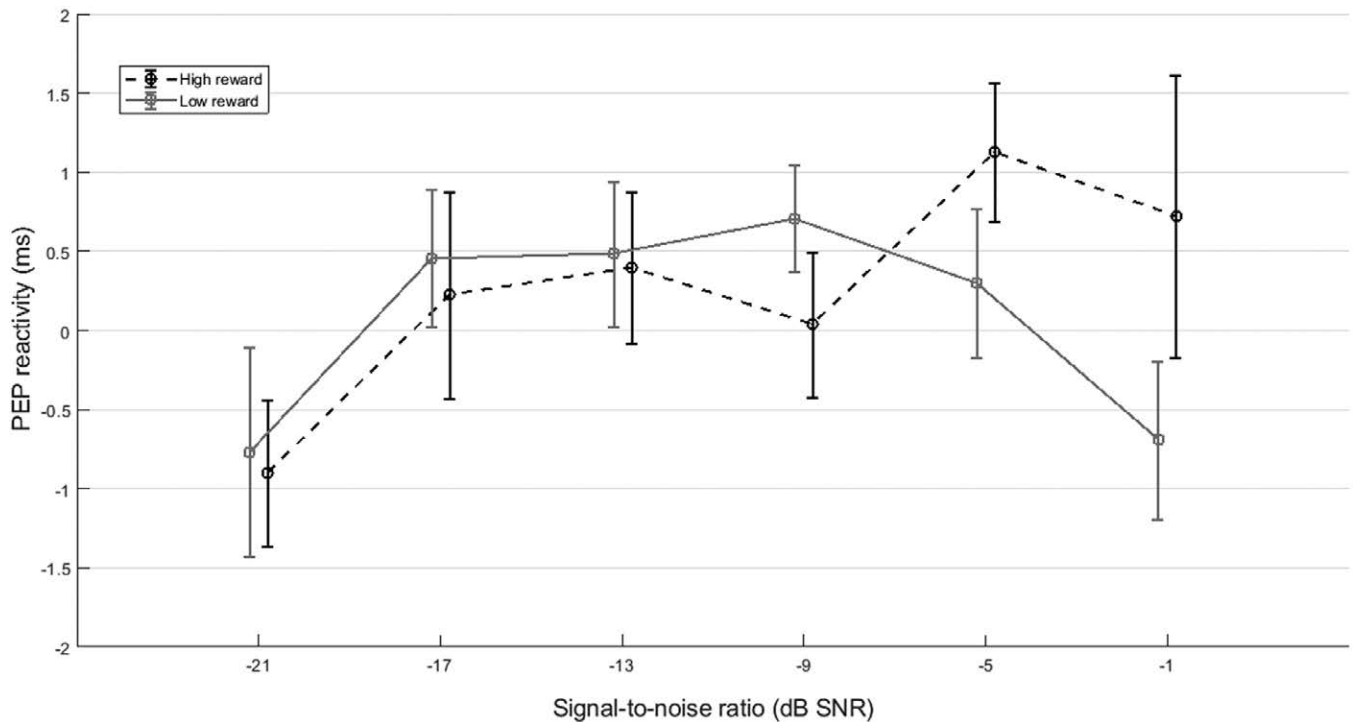
Fig. 4. Average PEP reactivity (in msec) was obtained using the block-wise analysis method. Error bars represent SEM. PEP, pre-ejection period.

comparison of the two hypotheses was completed for the block-wise PEP data.

Method 2: Target Stimuli Pre-ejection Period • The cycles occurring during the presentation of the target sentence were concatenated and analyzed. The average PEP reactivity scores across SNR are shown in Figure 5. In contrast to the block-wise PEP data discussed earlier, average PEP reactivity values for the target stimuli data were generally negative (the result of a shorter PEP during the task than the baseline period). The most negative PEP reactivity was detected at –21 dB SNR and the least negative reactivity was found at –5 dB SNR. A repeated measures ANOVA showed no statistically significant main effect of SNR ($F_{[3.49,104.70]}$ = 2.29, p = 0.07, ηp^2 = 0.07), reward ($F_{[1,30]}$ = 0.007, p = 0.94, ηp^2 < 0.001) or interaction between the two ($F_{[2.99,89.52]}$ = 0.69, p = 0.56, ηp^2 = 0.02). One-tailed quadratic and linear within-subject contrasts were conducted to test the relationship between SNR and PEP reactivity. The quadratic contrast revealed no significant effect of SNR ($F_{[.70,20.94]}$ = 2.86, p = 0.94, r = 0.09) whereas the linear contrast revealed a significant effect of SNR ($F_{[.70,20.94]}$ = 3.14, p = 0.05, r = 0.09), demonstrating a more negative PEP reactivity at lower SNRs. Likelihood analysis of the two hypotheses demonstrated that the data were 2.96 times more likely under the linear model than under the quadratic model.

Self-rating Scores

The first baseline video of each session was rated by the participants based upon how effortful and how difficult they found the video to watch. A score of 0 represented no effort or no difficulty, whereas 10 represented maximal effort or difficulty. Average baseline ratings of effort were 2.17 (SEM = 0.36) for the first session and 1.58 (SEM = 0.32) for the second session. Average difficulty ratings for the baseline period were

1.18 (SEM = 0.25) for the first session and 0.78 (SEM = 0.15) for the second session. Table 2 shows the average task-related self-rating scores and SEMs, and Table 3 shows the results of repeated measures ANOVAs for the self-rating scores. Self-rated effort investment, performance, difficulty, and giving up all demonstrated a significant effect of SNR (p < 0.001), but no effect of reward nor an interaction between reward and SNR. This showed that participants rated their effort investment, the task difficulty, and their tendency to give up as higher for lower SNRs. They rated their performance as lower for progressively lower SNRs.

Correlational Analysis: Reading Span Test and Need for Recovery Scale

Due to an error during RST data collection, one participant's data were excluded. RST data from 30 participants were subsequently included in the correlational analysis. The average RST score was 20.87 (SEM = 0.91) out of a total of 54. RST scores were not significantly correlated with any other variables, including average performance scores (r[26] = –0.56, p = 0.78), PEP reactivity (r[28] = –0.24, p = 0.19), and self-rated effort (r[28] = –0.13, p = 0.49), difficulty (r[28] = –0.59, p = 0.76), performance (r[28] = 0.05, p = 0.80), and tendency to give up (r[28] = –0.13, p = 0.50).

Regarding the NFR, one participant was excluded due to a missed answer during questionnaire completion. Therefore, 30 participants were included. The average NFR score was 38.79% (SEM = 4.36). Pearson's correlation coefficients revealed no significant correlations between the NFR scale and any of the other variables included (average performance score (r[26] = 0.18, p = 0.37), self-rated effort (r[28] = –0.34, p = 0.06), difficulty (r[28] = –0.20, p = 0.28), performance (r[28] = 0.21, p = 0.26), tendency to give up (r[28] = 0.06, p = 0.75), and the RST scores
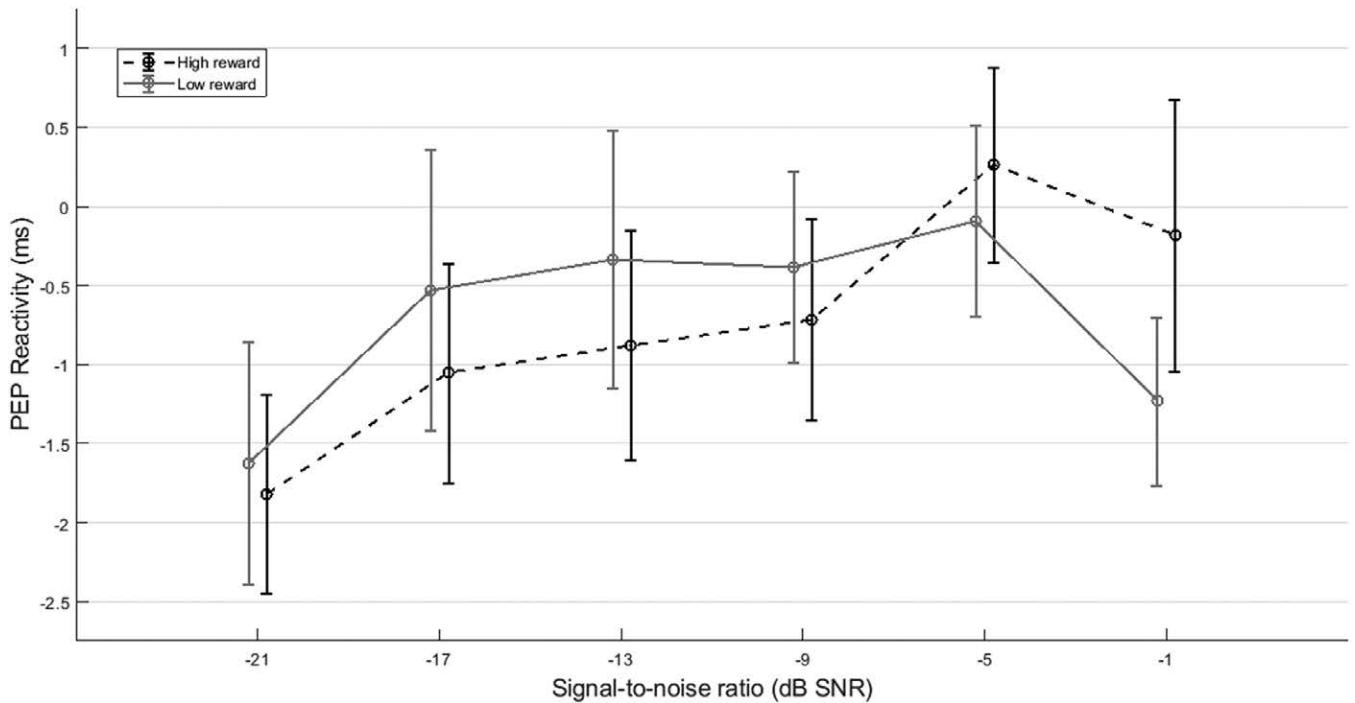
Fig. 5. Average PEP reactivity (in msec) was obtained using the target stimuli analysis method. Error bars represent SEM. PEP, pre-ejection period.

TABLE 2. Average self-rated effort, difficulty, performance, and tendency to give up

		dB SNR					
	Reward	−21	−17	−13	−9	−5	−1
SR effort	High	8.56 (0.18)	7.45 (0.15)	6.03 (0.17)	4.41 (0.29)	2.87 (0.26)	2.09 (0.26)
	Low	8.66 (0.15)	7.35 (0.17)	6.02 (0.26)	4.41 (0.24)	2.93 (0.24)	1.90 (0.24)
SR difficulty	High	8.26 (0.18)	7.00 (0.16)	5.58 (0.18)	3.71 (0.25)	2.30 (0.26)	1.56 (0.23)
	Low	8.37 (0.17)	7.00 (0.27)	5.50 (0.26)	3.54 (0.25)	2.13 (0.21)	1.14 (0.16)
SR performance	High	2.15 (0.15)	3.47 (0.19)	5.34 (0.22)	6.89 (0.26)	8.21 (0.17)	8.87 (0.18)
	Low	2.00 (0.18)	3.56 (0.21)	5.52 (0.24)	6.81 (0.21)	8.19 (0.19)	9.03 (0.15)
SR tendency to give up	High	6.05 (0.44)	4.65 (0.38)	3.17 (0.35)	2.08 (0.35)	1.35 (0.38)	0.94 (0.37)
	Low	6.01 (0.47)	4.56 (0.41)	3.14 (0.34)	2.37 (0.36)	1.79 (0.44)	1.08 (0.37)

The average score at each SNR for high- and low-reward conditions are presented. SEM is displayed in brackets. A higher self-rating score reflects more effort, increased difficulty, improved performance, and an increased tendency to give up.
dB SNR, decibel signal to noise ratio; SR, self-rated.

TABLE 3. Results of repeated measures analyses of variance for self-rating of effort, difficulty, performance and tendency to give up

	dB SNR				Reward				dB SNR × Reward			
	F	df	p	ηp²	F	df	p	ηp²	F	df	p	ηp²
SR effort	387.65	2.21, 66.35	**<0.001**	0.93	0.03	1, 30	0.86	<0.01	0.17	2.14, 64.10	0.86	<0.01
SR difficulty	432.37	2.67, 79.95	**<0.001**	0.94	1.52	1, 30	0.23	0.05	0.54	2.57, 77.03	0.63	0.02
SR performance	550.65	3.13, 93.75	**<0.001**	0.95	0.02	1, 30	0.88	<0.01	0.39	2.90, 87.08	0.75	0.01
SR tendency to give up	60.89	1.37, 40.94	**<0.001**	0.67	0.64	1, 30	0.43	0.02	0.61	3.35, 100.63	0.63	0.02

Where necessary, degrees of freedom and p values have been Greenhouse-Geisser corrected. p values reaching statistical significance are displayed in bold.
ηp², partial eta-square; SR, self-rated; SNR, signal to noise ratio.

(r[27] = 0.21, p = 0.26)). However, the Spearman rank revealed a significant moderate correlation between the NFR score and the average PEP reactivity (r[28] = −0.40, p = 0.03). This indicated that higher NFR scores (revealing greater need for recovery) were associated with more negative PEP reactivity values (more effort). A scatter plot of the relationship between NFR and average PEP reactivity is presented in Figure 6.

DISCUSSION

The main aim of this study was to demonstrate the relationship between motivation, task demand, and effort during a speech-in-noise task. This was achieved by varying reward levels across a range of SNRs while measuring PEP reactivity as an analog of effort. We also implemented two distinct methods
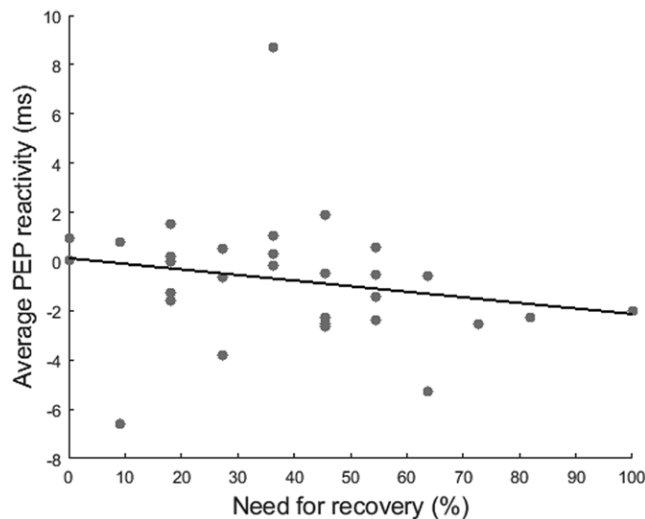
Fig. 6. Average PEP reactivity (in msec) (averaged across conditions) plotted against percentage need for recovery. Each point represents one individual participant and the solid line represents a best-fit regression line.

of PEP analysis: the block-wise method and the target stimuli method, with the aim of evaluating which is the most suitable for speech-recognition related PEP data. The final aspect of this study was to conduct a correlational analysis to examine the relationships between NFR and RST scores with average performance, self-ratings, and PEP reactivity.

Pre-ejection Period Analysis Methods

Two distinct methods of analyzing PEP data were utilized in this study. The *block-wise* method involved analyzing the whole block, including the masking noise presentation, target sentence, and participant response window. It is interesting that the majority of the block-wise average PEP reactivity values (Fig. 4) are positive, apparently showing that more effort was invested in the baseline than during the task. In contrast, the target stimuli method, which involved analyzing the cycles corresponding to the presentation of the target sentences only, showed generally negative average PEP reactivity values. These findings suggest that in speech-perception-type tasks, PEP reactivity is more sensitive to the task demand imposed by listening when specifically analyzing these data samples and removing data obtained from outside the target sentence presentation. It is likely that data obtained during the participant's vocalization, for example, adds noise to the signal and obscures the PEP response to listening. However, it is important to state that neither technique demonstrated statistically significant main effects of reward, SNR or any interaction between the two, and effect sizes were small for both analysis techniques. Possible reasons for this are explored later in this section after our specific hypotheses are discussed.

Pre-ejection Period Reactivity During Listening

Signal to Noise Ratio • We proposed two hypotheses regarding PEP reactivity and SNR. The first hypothesis stated that PEP reactivity would demonstrate a U-shaped relationship across SNR. This pattern implies that participants invest minimal effort in easy conditions, greatest effort at the difficult (but still possible) conditions, and less effort again at the most difficult conditions, due to disengagement. Our second hypothesis

was that PEP reactivity would be linearly related to SNR and disengagement would not be observed. The first hypothesis was tested with a quadratic contrast and the second was tested with a linear contrast.

For the block-wise PEP reactivity data, neither the quadratic nor linear models showed a statistically significant effect of SNR. Somewhat surprisingly, the pattern of the effort-task demand relationship appeared inverted with respect to our predictions. This would suggest that participants invested effort at the easiest and the most difficult conditions, with lower effort investment at the middle SNRs. Although no statistical significance was found, the displayed pattern contradicts both the presented hypotheses and also a body of work on objective measures of effort during speech-in-noise tasks (Zekveld & Kramer, 2014; Ohlenforst et al. 2018; Wendt et al. 2018; Zekveld et al. 2019). We suspect that the block-wise PEP reactivity scores reflect artefact and noise in the signal.

For the target stimuli PEP reactivity method, however, the Bayes factor revealed that the data best fit the linear, rather than the quadratic model. PEP reactivity varied linearly with SNR, such that participants invested more effort (as shown by a more negative PEP reactivity) as the task became progressively more challenging. It is interesting that no disengagement was seen at the more difficult SNRs (i.e., –21 dB SNR). It appears that participants did not perceive the task to be impossible, even though their sentence-based performance scores were very low. This finding was supported by the participants' self-rated tendency to give up: even at the most challenging condition (–21 dB SNR), the average tendency to give up rating was around 6.5 out of a total of 10, where higher scores indicate disengagement. This score is comparable to that previously presented by Zekveld and Kramer (2014). In their low intelligibility condition (0% correct), with a single-interfering-talker masking noise, the corresponding tendency to give up rating was 6.7 (Zekveld & Kramer, 2014).

The lack of disengagement at the lowest SNRs in our dataset could be due in part to three factors: (1) The type of masking noise presented during the task. A limitation of single-talker masking noise is that it is primarily dominated by informational masking, rather than energetic masking (Brungart, 2001). As such, the linguistic content of the masking speech increases the propensity for variability across trials and also confusion between the two speakers. In addition, the masking noise provided gaps free from both informational and energetic masking, during which it was possible to hear and understand odd words of the target stimulus, also known as dip listening (Wendt et al. 2018). The relatively high average word performance scores (compared to sentence scores) at the –21 dB SNR condition support the notion that participants were indeed successfully catching occasional words during the sentence. It is interesting that, although still low, the average number of nontarget words spoken at this difficulty level was higher than that at easier levels. It is possible that these additional words represent words substituted from the masking noise in error. This also provides weight to our suspicion that participants did not disengage.

(2) The instructions provided to the participants. Participants were encouraged to even repeat single words that they heard and were not informed that their performance results would be based upon whole sentences correctly recalled. It is likely that these instructions encouraged them to invest effort even when their performance was very low. (3) Finally, the presence of

the experimenter in close proximity during the task may have deterred the participants from giving up, even at the challenging conditions. It is known that effort investment related to social evaluative threat is sensitively shown by measures of the cardiovascular system (Woody et al. 2018). It seems that to induce disengagement as demonstrated by PEP reactivity, a more evenly distributed masking noise, such as a 4-talker babble, may have been more appropriate. Furthermore, it may have been beneficial to seat the experimenter outside the test room.

Reward

Our next hypothesis concerned the effect of reward level on PEP. We expected reward level to primarily affect effort investment at the lower SNRs. At these SNRs, we anticipated that higher reward would increase effort investment for the linear hypothesis or delay disengagement for the quadratic hypothesis. Yet no effect of reward was seen for either PEP analysis method. This finding was corroborated by a lack of any significant reward effect or Reward × SNR interaction in the self-rating data too. However, the performance data demonstrated a significant effect of reward, suggesting that participants did indeed find the high reward more motivating than the low reward. It is possible that the level of reward offered in this study (€0.20/5.00 per 5-minute-long block, over 12 blocks) was not enough to induce and measure a sensitive effect on PEP reactivity during such a lengthy test session. Similar studies using PEP have offered comparatively lower quantities of reward, but for a significantly shorter duration of the task (Richter, 2016). For instance, Richter's (2016) tone discrimination task offered the equivalent of around €0.19/1.90 per 3-minute-long block, of which there were just four. The reward levels in the present study were chosen to replicate those implemented by Koelewijn et al. (2018), who demonstrated an effect of reward on peak pupil dilation. In Koelewijn's study, participants undertook fewer intelligibility conditions divided into six task blocks, which in total lasted around an hour. It appears that in speech reception tasks pupil dilation may more sensitive to the effect of reward level than PEP. A larger reward level and a shorter session duration may have resulted in a reward main effect that could have been sensitively measured using PEP reactivity and demonstrated in the self-ratings.

Correlational Analysis with Reading Span Test and Need for Recovery

A correlational analysis was undertaken to explore whether there was a relationship between RST and NFR scores with average performance, self-ratings, and PEP reactivity. No significant correlations were found between the RST and any other variables. In particular, the lack of correlation between RST and performance was surprising, as working memory capacity has been widely demonstrated to relate to performance in speech-in-noise tasks (Koelewijn et al. 2012; Arehart et al. 2013; Desjardins & Doherty, 2013) and even the magnitude of the pupil response (Koelewijn et al. 2012; Wendt et al. 2016). The average RST score for our sample (20.87 (SD = 4.97) out of 54) was comparable to the average RST data reported from other young, normally hearing, native Dutch speakers (Besser et al. 2013). Our result suggests that in the sample of participants included in the present study, working memory capacity had no measurable relationship with performance, nor average PEP reactivity.

Our second expectation for the correlational analysis was to reveal a correlation between a person's NFR and their PEP reactivity. The results showed that those with a higher NFR score had a more negative average PEP reactivity, demonstrating that they invested more effort in general during the listening task. This contradicts findings of previous work which revealed that people with a higher NFR had a smaller peak pupil dilation during a speech-in-noise task (Wang et al. 2018). An explanation for these contrasting findings is that the influence of fatigue on these two distinct physiological measures may differ. Particularly as the pupil response reflects a combination of both sympathetic and parasympathetic nervous system activity, whereas PEP is solely related to sympathetic nervous system activity. Generally, studies using cardiovascular measures during nonlistening tasks have shown that provided the task is judged to be possible, fatigue leads to a greater effort investment (Schmidt et al. 2010; Mlynski et al. 2017). The rationale, based upon MIT, is that fatigued individuals have depleted resources and therefore perceive easier tasks to be more demanding than their nonfatigued counterparts do (Wright et al. 2008). As a result, those with fatigue are thought to invest compensatory effort when facing a challenge (Hockey, 1997). It is expected that fatigued individuals will disengage at a lower demand level than their nonfatigued peers, due to reduced perception of their abilities and a depleted supply of resources. In the present study, disengagement was not seen in the PEP reactivity data, which suggests that participants still considered the task to be possible. As such, it is plausible that those in the study who reported a higher need for recovery invested more effort than those with a lower need for recovery.

Limitations of the Study

As demonstrated earlier, PEP reactivity changes in response to a speech-in-noise paradigm were modest. This could relate to several methodological limitations of the study. For example, one limitation of the present study relates to the timings of the used PEP recording paradigm during listening. In short, sentence-based speech perception tasks, the peak pupil dilation is normally measured temporally at the end of the target stimulus presentation, during several seconds of masking noise that occurs before the participant's response. This time period has been referred to as a "retention phase" (Winn et al. 2018). A retention phase was missing from the design of the present study. It is possible that measuring PEP during the target stimulus fails to capture aspects of listening effort that occur temporally later. Future work should implement a retention phase in the design, such that the optimal time for recording the PEP response during a speech-in-noise task can be determined. Another consideration for future studies is the duration of speech stimulus presentation. In a traditional speech-in-noise paradigm, such as that used in this work, the nature of the listening involved is short and staccato, punctuated by nontarget listening and vocalization. Presenting a narrative, rather than short sentences as in the present study, would increase the duration of active listening, which might in turn increase the sensitivity of the measure. Lastly, it was beyond the scope of this study to implement measures of HRV in addition to the presented work. However, it may be prudent to include both PEP and high-frequency HRV in future work, to establish and compare the relative contributions of the sympathetic and parasympathetic cardiac responses during listening.

CONCLUSION

This study measured PEP reactivity for the first time during a speech-in-noise task. Of the two used data analysis methods, the target stimuli method was determined to be superior, showing a more prominent PEP response than the block-wise method. Predicted reward effects were absent in the present dataset and no evidence of disengagement was demonstrated. Instead, PEP reactivity varied linearly with task demand alone. Finally, participants with a higher need for recovery invested more overall effort, as displayed by a more negative average PEP reactivity.

REFERENCES

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

Arehart, K. H., Souza, P., Baca, R., Kates, J. M. (2013). Working memory, age, and hearing loss: susceptibility to hearing aid distortion. *Ear Hear,* 34, 251–260.

Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., Festen, J. M. (2013). How linguistic closure and verbal working memory relate to speech recognition in noise–a review. *Trends Amplif,* 17, 75–93.

Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annu Rev Psychol,* 40, 109–131.

Brungart, D. S. (2001). Evaluation of speech intelligibility with the coordinate response measure. *J Acoust Soc Am,* 109(5 Pt 1), 2276–2279.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *J Verb Learn Verb Behav,* 19, 450–466. https://doi.org/10.1016/s0022-5371(80)90312-6

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

Glover, S., & Dixon, P. (2004). Likelihood ratios: a simple and flexible statistic for empirical psychologists. *Psychon Bull Rev,* 11, 791–806.

Hällgren, M., Larsby, B., Lyxell, B., Arlinger, S. (2001). Evaluation of a cognitive test battery in young and elderly normal-hearing and hearing-impaired persons. *J Am Acad Audiol,* 12, 357–370.

Hockey, G. R. J. (1997). Compensatory control in the regulation of human performance under stress and high workload: A cognitive-energetical framework. *Biol Psychol,* 45, 73–93. https://doi.org/10.1016/S0301-0511(96)05223-4

Holman, J. A., Drummond, A., Hughes, S. E., Naylor, G. (2019). Hearing impairment and daily-life fatigue: a qualitative study. *Int J Audiol,* 58, 408–416.

Hopstaken, J. F., van der Linden, D., Bakker, A. B., Kompier, M. A. (2015). The window of my eyes: Task disengagement and mental fatigue covary with pupil dynamics. *Biol Psychol,* 110, 100–106.

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

Jennings, J. R., Kamarck, T., Stewart, C., Eddy, M., Johnson, P. (1992). Alternate cardiovascular baseline assessment techniques: vanilla or resting baseline. *Psychophysiology,* 29, 742–750.

Kahneman, D. (1973). *Attention and Effort Englewood Cliffs,* New Jersey.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *Int J Otolaryngol,* 2012, 865731.

Koelewijn, T., Zekveld, A. A., Lunner, T., Kramer, S. E. (2018). The effect of reward on listening effort as reflected by the pupil dilation response. *Hear Res,* 367, 106–112.

Krohova, J., Czippelova, B., Turianikova, Z., Lazarova, Z., Tonhajzerova, I., Javorka, M. (2017). Preejection period as a sympathetic activity index: a role of confounding factors. *Physiol Res,* 66(Suppl 2), S265–S275.

Loewenfeld, I. E., & Lowenstein, O. (1999). *The pupil : anatomy, physiology, and clinical applications.* Butterworth-Heinemann.

Mackersie, C. L., & Calderon-Moultrie, N. (2016). Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear Hear,* 37 Suppl 1, 118S–125S.

Mackersie, C. L., MacPhee, I. X., Heldt, E. W. (2015). Effects of hearing loss on heart rate variability and skin conductance measured during sentence recognition in noise. *Ear Hear,* 36, 145–154.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behav Res Methods,* 43, 679–690. https://doi.org/10.3758/s13428-010-0049-5

Mazeres, F., Brinkmann, K., & Richter, M. (2019). Implicit achievement motive limits the impact of task difficulty on effort-related cardiovascular response. *J Res Pers,* 82, 103842. https://doi.org/10.1016/j.jrp.2019.06.012

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology,* 54, 193–203.

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'. *Int J Audiol,* 53, 433–440.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol Methods,* 1(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Mlynski, C., Wright, R. A., Agtarap, S. D., Rojas, J. (2017). Naturally-occurring fatigue and cardiovascular response to a simple memory challenge. *Int J Psychophysiol,* 119, 73–78.

Newlin, D. B., & Levenson, R. W. (1979). Pre-ejection period: measuring beta-adrenergic influences upon the heart. *Psychophysiology,* 16, 546–553.

Obrist, P. A., Light, K. C., James, S. A., Strogatz, D. S. (1987). Cardiovascular responses to stress: I. Measures of myocardial response and relationship to high resting systolic pressure and parental hypertension. *Psychophysiology,* 24, 65–78.

Ohlenforst, B., Wendt, D., Kramer, S. E., Naylor, G., Zekveld, A. A., Lunner, T. (2018). Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. *Hear Res,* 365, 90–99.

Ohlenforst, B., Souza, P. E., MacDonald, E. N. (2017a). Response to Comment: RE: Exploring the relationship between working memory, compressor speed, and background noise characteristics, *Ear Hear* 37, 137–143. *Ear Hear,* 38, 644–645.

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., Kramer, S. E. (2017b). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hear Res,* 351, 68–79.

Pan, J., & Tompkins, W. J. (1985). A real-time QRS detection algorithm. *IEEE Trans Biomed Eng,* 32, 230–236.

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

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 Hear,* 37, 5S–27S.

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *J Speech Hear Res,* 29, 146–154.

Richter, M. (2016). The moderating effect of success importance on the relationship between listening demand and listening effort. *Ear Hear,* 37 Suppl 1, 111S–117S.

Richter, M., Friedrich, A., Gendolla, G. H. (2008). Task difficulty effects on cardiac activity. *Psychophysiology,* 45, 869–875.

Richter, M., & Gendolla, G. H. (2009a). The heart contracts to reward: monetary incentives and preejection period. *Psychophysiology,* 46, 451–457.

Richter, M., & Gendolla, G. H. E. (2009b). Mood impact on cardiovascular reactivity when task difficulty is unclear. *Motivation and Emotion*, 33, 239–248. https://doi.org/10.1007/s11031-009-9134-4

Richter, M., & Knappe, K. (2014). Mood impact on effort-related cardiovascular reactivity depends on task context: evidence from a task with an unfixed performance standard. *Int J Psychophysiol,* 93, 227–234.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front Syst Neurosci, JUNE*. https://doi.org/10.3389/fnsys.2013.00031

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *Int J Audiol*, 47(SUPPL. 2). https://doi.org/10.1080/14992020802301167

Schmidt, R. E., Richter, M., Gendolla, G. H., Van der Linden, M. (2010). Young poor sleepers mobilize extra effort in an easy memory task: evidence from cardiovascular measures. *J Sleep Res,* 19, 487–495.

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *J Speech Lang Hear Res,* 58, 1781–1792.

Sherwood, A., Allen, M. T., Fahrenberg, J., Kelsey, R. M., Lovallo, W. R., van Doornen, L. J. (1990). Methodological guidelines for impedance cardiography. *Psychophysiology,* 27, 1–23.

Sherwood, A., Allen, M. T., Obrist, P. A., Langer, A. W. (1986). Evaluation of beta-adrenergic influences on cardiovascular and metabolic adjustments to physical and psychological stress. *Psychophysiology,* 23, 89–104.

Steinhauer, S. R., Siegle, G. J., Condray, R., Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *Int J Psychophysiol,* 52, 77–86.

Van Veldhoven, M., & Broersen, S. (2003). Measurement quality and validity of the "need for recovery scale." *Occup Environ Med,* 60(Suppl. 1), i3.

Versfeld, N. J., Daalder, L., Festen, J. M., Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J Acoust Soc Am,* 107, 1671–1684.

Wang, Y., Naylor, G., Kramer, S. E., Zekveld, A. A., Wendt, D., Ohlenforst, B., Lunner, T. (2018a). Relations between self-reported daily-life fatigue, hearing status, and pupil dilation during a speech perception in noise task. *Ear Hear,* 39, 573–582.

Wang, Y., Metz, J., Costello, J. L., Passmore, J., Schrader, M., Schultz, C., Islinger, M. (2018b). Intracellular redistribution of neuronal peroxisomes in response to ACBD5 expression. *PLoS One,* 13, e0209507.

Wendt, D., Dau, T., Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Front Psychol,* 7, 345.

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hear Res,* 369, 67–78.

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 Hear,* 22, 2331216518800869.

Woody, A., Hooker, E. D., Zoccola, P. M., Dickerson, S. S. (2018). Social-evaluative threat, cognitive load, and the cortisol and cardiovascular stress response. *Psychoneuroendocrinology,* 97, 149–155.

Wright, R. A. (1996). Brehm's theory of motivation as a model of effort and cardiovascular response. *The Psychology of Action: Linking Cognition and Motivation to Behavior*, *January 1996*, 424–453. https://www.researchgate.net/publication/280646002_Brehm's_theory_of_motivation_as_a_model_of_effort_and_cardiovascular_response

Wright, R. A., Stewart, C. C., Barnett, B. R. (2008). Mental fatigue influence on effort-related cardiovascular response: extension across the regulatory (inhibitory)/non-regulatory performance dimension. *Int J Psychophysiol,* 69, 127–133.

Wu, Y. H., Stangl, E., Zhang, X., Perkins, J., Eilers, E. (2016). Psychometric functions of dual-task paradigms for measuring listening effort. *Ear Hear,* 37, 660–670.

Zekveld, A. A., Koelewijn, T., Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: current state of knowledge. *Trends Hear,* 22, 2331216518777174.

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

Zekveld, A. A., van Scheepen, J. A. M., Versfeld, N. J., Veerman, E. C. I., Kramer, S. E. (2019). Please try harder! The influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels. *Hear Res,* 381, 107768.

Zhang, M., Siegle, G. J., McNeil, M. R., Pratt, S. R., Palmer, C. (2019). The role of reward and task demand in value-based strategic allocation of auditory comprehension effort. *Hear Res,* 381, 107775.