



## LJMU Research Online

**Plain, BJ, Pielage, H, Zekveld, AA, Richter, M, Bhuiyan, TA, van de Ven, SRB and Kramer, SE**

**Incorporating virtual reality agents during a dichotic speech reception task: insights from the heart**

<http://researchonline.ljmu.ac.uk/id/eprint/24548/>

### Article

**Citation** (please note it is advisable to refer to the publisher's version if you intend to cite from this work)

**Plain, BJ, Pielage, H, Zekveld, AA, Richter, M, Bhuiyan, TA, van de Ven, SRB and Kramer, SE (2024) Incorporating virtual reality agents during a dichotic speech reception task: insights from the heart. Ear and Hearing, 46 (2). pp. 444-460. ISSN 0196-0202**

LJMU has developed [LJMU Research Online](#) for users to access the research output of the University more effectively. Copyright © and Moral Rights for the papers on this site are retained by the individual authors and/or other copyright owners. Users may download and/or print one copy of any article(s) in LJMU Research Online to facilitate their private study or for non-commercial research. You may not engage in further distribution of the material or use it for any profit-making activities or any commercial gain.

The version presented here may differ from the published version or from the version of the record. Please see the repository URL above for details on accessing the published version and note that access may require a subscription.

For more information please contact [researchonline@ljmu.ac.uk](mailto:researchonline@ljmu.ac.uk)

<http://researchonline.ljmu.ac.uk/>

# Incorporating Virtual Reality Agents During a Dichotic Speech Reception Task: Insights From the Heart

Bethany Plain,<sup>1,2</sup> Hidde Pielage,<sup>1,2</sup> Adriana A. Zekveld,<sup>1</sup> Michael Richter,<sup>3</sup> Tanveer A. Bhuiyan,<sup>4</sup> Sjors R. B. van de Ven,<sup>1</sup> and Sophia E. Kramer<sup>1</sup>

**Objectives:** Listening effort is moderated by not only task difficulty, but also success importance. In real communication scenarios, success importance varies based upon the social context. However, in the laboratory, it can be challenging to manipulate social context without compromising experimental control. Outside of hearing sciences, studies have applied virtual reality (VR) to incorporate social context in a controlled and repeatable manner. Several of these studies have demonstrated that social manipulations in VR can reliably elicit changes in cardiovascular measures. Here, we investigated the effect of adding VR agents to a speech reception task, while measuring from the cardiovascular system.

**Design:** Twenty-eight, normally hearing participants undertook a dichotic speech reception task. Sentences in stationary noise were presented dichotically, that is, different sentences presented simultaneously to each ear. Participants were tasked to either repeat one of the sentences (single-sentence condition) or both of the sentences (dual-sentence condition). The task was conducted under two VR conditions: (1) in the presence of agents, who provided sporadic performance feedback and (2) in the presence of nonagent controls, without any feedback given. Alongside task performance, we quantified changes in cardiovascular measures, relative to pretask baselines: heart rate variability, pre-ejection period, heart rate, and blood pressure. After each condition, participants rated their subjective effort, difficulty, performance, and engagement.

**Results:** Performance and the subjective perception of performance were lower, while subjective effort and difficulty were higher, in the dual-sentence condition, compared with the single-sentence condition. Heart rate was the only cardiovascular measure that was sensitive to the experimental manipulations. Contrary to our expectations, heart rate increased in the nonagent control conditions, compared with the agent conditions. An exploratory analysis revealed heart rate fluctuations within a trial: heart rate was higher during the first 6 sec of the trial (reflecting the presentation masking noise and the sentence presentation) in the dual-sentence condition, compared with the single-sentence condition.

**Conclusions:** This study was the first to incorporate VR agents who provided performance feedback during a dichotic speech reception task. Our results suggest that the VR agents did not increase success importance, which could be attributed to a lack of realism of the agents. We also demonstrated that the cardiovascular response to experimental manipulations may differ depending on the data window selected for analysis.

**Key words:** Autonomic nervous system, Cardiovascular, Dichotic listening, Listening effort, Virtual reality.

(*Ear & Hearing* 2025;46:444–460)

## INTRODUCTION

Listening effort is an increasingly investigated phenomenon, defined as “the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task” (Pichora-Fuller et al. 2016, p. 11S). In daily life, speech is commonly encountered in the presence of acoustic challenges (or obstacles), for example, interfering background noises, which must be ignored to selectively attend to the target talker (Smeds et al. 2015; Wu et al. 2018; Christensen et al. 2021; Shields et al. 2022). The process of directing attention in this manner can be demanding and can require substantial listening effort investment (Pichora-Fuller et al. 2016; Koelewijn et al. 2017). The increase in effort required when attention is divided between different speech sources has been demonstrated in several laboratory studies (Best et al. 2010; Koelewijn et al. 2014; Seeman & Sims 2015).

Another important aspect affecting listening effort in daily communication scenarios is the social context. Social context is thought to moderate “success importance”: the value that the listener places on understanding successfully (Matthen 2016; Pichora-Fuller 2016; Pichora-Fuller et al. 2016; Hughes et al. 2018; Shields et al. 2022). Considering the social contexts at play, a listener may determine whether the required effort is justified, and if not, they may disengage from listening (Brehm & Self 1989; Mackersie & Kearney 2017). When investigating listening effort in the laboratory, however, it is challenging to manipulate social context without compromising experimental control (Kvavilashvili & Ellis 2004). One way to maintain the desired level of experimental control and consistency is the application of virtual reality (VR) technology. In this study, we investigated the effect of incorporating VR agents during a dichotic speech reception task.

## Dichotic Listening

Attention is “a multidimensional construct that includes orienting, selecting, and/or focusing on environmental stimuli (e.g., speech) or internal representations (e.g., thoughts)” (Pichora-Fuller et al. 2016, p. 11S). Several studies have investigated the effects of dichotic listening in the laboratory (Best et al. 2006, 2010; Shinn-Cunningham & Best 2008; Koelewijn et al. 2014). For example, Best et al. (2010) presented simultaneous, independent streams of masked speech dichotically to normal-hearing listeners. Participants undertook two different task conditions: a single-attention condition, where they had to attend to and repeat key words

<sup>1</sup>Amsterdam University Medical Center, Vrije Universiteit Amsterdam, Otolaryngology Head and Neck Surgery, Ear & Hearing, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands; <sup>2</sup>Eriksholm Research Centre, Snekkersten, Denmark; <sup>3</sup>School of Psychology, Faculty of Health, Liverpool John Moores University, Liverpool, United Kingdom; and <sup>4</sup>Demant A/S, Kongebakken, Smørum, Denmark.

Copyright © 2024 The Authors. *Ear & Hearing* is published on behalf of the American Auditory Society, by Wolters Kluwer Health, Inc. This is an open access article distributed under the Creative Commons Attribution License 4.0 (CCBY), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

presented to one ear, while ignoring stimuli simultaneously presented to the other ear, or a dual-attention condition, where they had to attend to and repeat key words presented simultaneously to both ears. In the dual-attention condition, the accuracy at which participants could repeat target words deteriorated, compared with the single-attention condition (Best et al. 2010).

Physiological measures have been used during dichotic listening tasks, as a proxy of listening effort (Koelewijn et al. 2014; Seeman & Sims 2015). For example, Koelewijn et al. (2014) conducted a similar study design to that of Best et al. (2010), with the addition of pupil dilation measures. They reported that the dual-attention condition elicited increased cognitive processing load (demonstrated by a larger pupil dilation) as compared with the single-attention condition. Rather than pupil measures, Seeman and Sims (2015) incorporated cardiovascular measures into their dichotic listening task. Their participants repeated digits in three different configurations, increasing in task complexity: diotic single digit (one digit to repeat), dichotic single digits (two digits to repeat, one from each ear) and dichotic double digits (four digits to repeat, two from each ear). The results demonstrated a significantly larger change in heart rate variability (HRV) relative to baseline during the two dichotic conditions, compared with the diotic condition, suggesting increased effort investment (Seeman & Sims 2015).

### Social Context and VR

The social context of the communication scenario is also important in determining listening effort investment (Matthen 2016; Pichora-Fuller 2016; Hughes et al. 2018). Despite this knowledge, laboratory research into listening effort has rarely implemented social context manipulations. To our knowledge, just five studies have investigated social context during speech-in-noise tasks, each applying a different paradigm: (1) participants answered a set of questions evaluating the level of their comprehension of the speech material they heard (Picou & Ricketts 2014), (2) participants were told that video footage of them performing the task would be later evaluated by a panel of experts (Mackersie & Kearney 2017), (3) pairs of participants performed the task in tandem (Pielage et al. 2021), (4) two strangers passively observed participants during a listening task (Plain et al. 2021; Pielage et al. 2023), and (5) the experimenter provided explicit performance feedback and encouragement to “please try harder” during the task (Zekveld et al. 2019). The evaluative manipulations in these studies aimed to draw upon social-evaluative threat, or the fear of being negatively judged by others, which is commonly experienced in daily life (Dickerson & Kemeny 2004; Jonas & Mühlberger 2017).

These five paradigms varied in their level of experimental control and strength of evaluation, as well as their degree of ecological validity (Keidser et al. 2020). The evaluative aspect of the first two studies was achieved by means of the instructions provided to participants. In their evaluated condition, Picou and Ricketts’ (2014) participants were told that their comprehension of the sentence material would later be evaluated by means of a set of questions. In Mackersie and Kearney (2017) study, participants were told they were being video recorded and that a panel of experts would later review the footage. While providing evaluation by such means is well-controlled and highly repeatable, studies investigating cortisol responses

to social-evaluative threat have revealed that video recording-based methods elicit smaller responses than evaluation by physically present evaluative others (Dickerson & Kemeny 2004). It is likely that real-time evaluation and the physical presence of an evaluator are more relevant to the participant than video footage of them being reviewed later.

The evaluative aspect of the other three studies included the physical presence of one or more human interlocutors, with the designs varying depending on whether these individuals were untrained or trained. Two of the studies (2019) included non-trained individuals in their designs. Pielage et al. (2021) included two participants in the same room simultaneously, who both performed speech perception testing by alternating sentence repetition during the task. In contrast, Plain et al. and Pielage et al. (2023) included two strangers who were seated within the participant’s eyeline and passively observed and evaluated them performing the task. It could be argued that these two experiments were susceptible to variability and lacked experimental control, because people are inherently unpredictable in their behavior. For example, even though interactions between the individuals were minimal, each participant pair or trio will have had a slightly different social dynamic. Such social interactions are difficult to control, quantify, and replicate.

These limitations affecting experimental control can be alleviated by the inclusion of evaluators trained using a standardized protocol. Zekveld et al.’s (2019) participants were explicitly urged by the experimenter—a trained individual—to try harder to improve their performance to reach an impossible target performance level. Here, the evaluation was repeatable and standardized across study participants. Outside of hearing sciences, the impact of trained evaluators is well documented, particularly in the literature regarding the Trier Social Stress Test (TSST). The TSST involves participants undertaking challenging tasks, including public speaking and mental arithmetic, in the presence of a physical audience (Dickerson & Kemeny 2004). This evokes physiological responses, including neuroendocrine and cardiovascular responses, interpreted to reflect social-evaluative threat (Allen et al. 2014). However, it has been noted that the TSST procedure may vary slightly between sites due to different laboratory environments and different audiences (Jönsson et al. 2010). In addition, it is resource-heavy due to the requirement for trained evaluators (Shiban et al. 2016).

A promising way in which to simulate social scenarios in the laboratory, while maintaining experimental control and repeatability, is the development of VR test scenarios (Kothgassner & Felnhofer 2020). The impact of a VR TSST has been investigated and crucially, VR adaptations of the TSST (i.e., the audience members are agents rather than physical people), have been shown to elicit similar, albeit in some cases smaller, physiological responses to the real-life TSST (Zimmer et al. 2019; Fallon et al. 2021). To our knowledge, no previous listening effort study has included agents in VR to provide performance evaluation.

### Cardiovascular Measures During Listening Tasks

During listening tasks, physiological responses have been reported as correlates of listening effort, including changes to pupil size, skin conductance, and cardiovascular measures (McGarrigle et al. 2014; Pichora-Fuller et al. 2016; Zekveld et al. 2018). Though useful indicators, it should be noted that the

physiological measures are unspecific: alone, they do not enable distinction between different cognitive processes (Zekveld et al. 2018; Richter et al. 2023). For example, effort and emotion may result in similar physiological responses (Francis & Love 2020). Of the aforementioned measures, cardiovascular measures are relatively unique as they allow for simultaneous and distinct assessment of activity from both branches of the autonomic nervous system: the sympathetic nervous system, SNS, and the parasympathetic nervous system, PNS (Berntson et al. 1991; Lovallo 2005; McCorry 2007). For example, pre-ejection period (PEP), the interval between the onset of excitation of the left ventricle and the opening of the aortic valve, is predominantly a measure of SNS activity (Ahmed et al. 1972; Newlin & Levenson 1979), whereas HRV, representing the fluctuation of the interval between heart beats, is considered a measure of PNS activity, depending on the metric used (Shaffer & Ginsberg 2017). Other relevant psychophysiological measures including heart rate and blood pressure are controlled by mixed autonomic origins (Gordan et al. 2015; Plain et al. 2021).

Due to their autonomic origins, cardiovascular measures hold promise in demonstrating physiological changes associated with effort. Despite this, the current picture of cardiovascular reactivity during listening is somewhat unclear. During speech reception tasks, HRV has been demonstrated to be sensitive to changes in talker rate (Mackersie & Calderon-Moultrie 2016) and signal to noise ratio (SNR) (Mackersie et al. 2015; Seeman & Sims 2015), but similar effects were notably absent in other studies (Mackersie et al. 2015; Plain et al. 2021). PEP has also been shown to be sensitive to SNR in some studies (Plain et al. 2020; Slade et al. 2021), but not in others (Plain et al. 2021).

One possible reason for these inconsistent results may relate to the data window used for the analysis. In general, studies have averaged measures across a whole task block, including data collected during masking noise presentation, sentence presentation, the participant verbal response, and experimenter scoring time (Mackersie et al. 2015; Seeman & Sims 2015; Mackersie & Calderon-Moultrie 2016; Plain et al. 2020, 2021). It is possible that by doing so, some of the dynamics of transient responses at the trial level are overlooked. Supporting this, a recent study demonstrated a more sensitive PEP response to SNR when the data corresponding to the stimulus presentation interval were analyzed only, as opposed to all data within the block (Plain et al. 2020). Furthermore, Francis et al. (2016) demonstrated that heart rate fluctuated during trials of a speech reception task. They showed that compared with a pretrial baseline, heart rate decreased at around 6 sec after trial onset. More work is needed to elucidate the trial-level dynamics of cardiovascular measures.

### Aims and Hypotheses

The present study manipulated task demand during a dichotic listening task, while incorporating agents or nonagent controls (stacked boxes displaying small idle movements; see more detail in VR manipulation: agents and nonagent controls) into the virtual environment. The task consisted of two dichotic listening conditions: single- and dual-sentence. In both, speech stimuli were presented dichotically (i.e., different sentences presented to each ear in stationary noise). In the single-sentence condition, participants repeated the sentence presented to one ear, ignoring the sentence in the other ear, whereas in the dual-sentence condition, they repeated both sentences, one presented

to each ear. We manipulated success importance by adding agents or nonagent controls into the virtual environment: participants performed the task in the presence of two agents who provided occasional nonverbal performance feedback (head nod for correct answers, head shake for incorrect answers), or in front of nonagent visual controls without feedback. We measured sentence repetition performance, self-reported measures of effort, task difficulty, performance and engagement, and physiological responses from the cardiovascular system and pupils of participants (the pupil data will be presented elsewhere).

We hypothesized that sentence repetition performance (proportion of words from both sentences) would decrease in the dual-sentence condition compared with the single-sentence condition, as demonstrated in similar studies (Best et al. 2010; Koelewijn et al. 2014). We expected that in the single-sentence condition, the presence of agents and feedback would increase success importance, motivating participants to invest more effort compared with the nonagent control condition. This would be reflected by the self-report measures and indexed by a shortening of PEP, a decrease in high-frequency HRV (HF-HRV) and an increase in heart rate. In the dual-sentence condition, we expected an overall increase in effort investment compared with the single-sentence condition, because the task was more challenging. We also anticipated that the effect of the virtual agents would be diminished in the dual-sentence condition because the cognitive load of dividing attention would override the effect of the agents. That is, the dual-sentence condition may be demanding enough to elicit a ceiling effect, such that the presence of the agents would have no measurable additional impact.

An additional aim of the study was to explore the timings of cardiovascular changes within speech reception task trials. There is evidence to suggest that changes to heart rate occur relatively quickly, that is, within seconds, in response to sound (Francis et al. 2016; Shoushtarian et al. 2019). Furthermore, as described earlier, a previous study revealed higher sensitivity of PEP reactivity when only the listening components were analyzed (Plain et al. 2020). By averaging across full blocks of sentences it is possible that physiological changes are missed. In this additional exploration, we tested the same hypotheses but instead selected three parts of the trial to analyze the physiological responses in the following intervals: during the masking noise preceding the sentence, during sentence presentation, and finally, during the postsentence masking noise.

## MATERIALS AND METHODS

### Participants

The study was advertised by means of distribution of flyers, both physically at university buildings and via Facebook posts shared in local groups. To determine the sample size, a power calculation was performed upon previous work using HRV ( $0.74 \eta_p^2$ ) and heart rate ( $0.25 \eta_p^2$ ) during different task complexity levels of a dichotic digit repetition task (Seeman & Sims 2015). Our power calculation, completed in G\*Power 3.1.9.4 software, referred to a repeated measures analysis of variance (ANOVA) with an estimated univariate effect size of 0.25, an alpha error of 0.05, power of 0.8, and a correlation of 0.5 between repeated measures. This calculation indicated that 24 participants were required. Four more were included, as we anticipated a high possibility of data exclusion when preprocessing the physiological measures.



First, 12 pilot participants were recruited and tested with a shortened experimental protocol without baselines. Subsequently, 28 participants underwent the full experimental protocol—results from these participants are reported here. Two of these had to be excluded because of cardiovascular data quality issues, therefore data from the remaining 26 participants (11 males, 15 females) will be presented here. Participants were native Dutch speakers, right-handed, and normally hearing ( $\leq 20$  dB HL at 0.5, 1, 2, and 4 kHz), with a mean age of 26.5 years ( $SD = 4.4$  years). They reported no history of psychiatric, neurological, ocular, or cardiovascular problems. Participants with long or short-sightedness were encouraged to wear lenses, rather than glasses, during the experiment. All participants provided informed consent in accordance with local ethical committee procedures.

### Procedure and Apparatus

Testing was conducted in a sound-treated room. Participants were seated on a centrally placed chair, in front of a table. The experimenter controlled the equipment from an adjacent room, outside the participant's view. During the experiment, participants wore an HTC Vive pro VR headset with headphone attachments, which were used to present all audio stimuli. The virtual environment was designed and implemented in 3D game engine software, unity, using additional SteamVR software. Custom-made C# scripts were written to run the experimental protocol.

Participants attended a single test session lasting around 2 hr in duration. At the start of the test session, the experimenter explained all procedures and obtained the participant's written informed consent to proceed. Following this, pure tone audiometry was conducted at 0.5, 1, 2, 4, and 8 kHz, to ensure that participants met the audiometric inclusion criteria. Next, participants had their height and weight measured by the experimenter, such that body mass index (BMI) could be calculated. Then, the experimenter applied the cardiovascular electrodes and blood pressure cuff. After a thorough explanation of the task, the VR headset was applied to the participant and calibration of the eye tracker within the VR headset was conducted.

Participants were then given the opportunity to acclimate to the VR environment, while they practiced the task: four practice trials were presented, two single-sentence trials, and two dual-sentence trials. During the practice, participants were also familiarized with the agents, nonagent controls, and the animations of both. Aside from moving their head and observing the animated objects, participants could not interact with the VR environment. Subsequently, the four task blocks were conducted. Each task block contained 30 sentences and lasted around 10 min in total. The order of task conditions was counterbalanced between participants, resulting in 24 unique order combinations with two repeated orders.\* Each block was preceded by a 3-min baseline video and ended with the participants removing the headset to complete subjective rating scales on

paper. Electrocardiography (ECG) and impedance cardiography (ICG) were measured throughout the experiment. The blood pressure cuff was inflated once in the middle of each baseline and once during each task block. Pupil size was also recorded during the task blocks. The pupil data will be reported elsewhere (Pielage et al. n.d.). Breaks were offered to participants after two blocks. When the four task blocks had been completed, participants completed two questionnaires: the i-group presence questionnaire (Schubert et al. 2001) and a questionnaire about the agents. Last, participants were debriefed and the procedure for reimbursement was discussed.

### Dichotic Listening Test Blocks

The experiment consisted of a two (dichotic speech reception task demand) by two (VR manipulation) within-subject design.

**Dichotic Speech Reception Task Demand** • The dichotic speech reception task was inspired by previous work (Shinn-Cunningham & Best 2008; Best et al. 2010; Koelewijn et al. 2014). Per task block, 30 Dutch, everyday sentences were presented dichotically in speech-shaped stationary noise, the level of which was 65 dB SPL (Versfeld et al. 2000). The stationary noise, referred to here as masking noise, preceded each sentence by 3 sec and continued for 3 sec after sentence offset. Sentence presentation lasted on average 1.84 sec but ranged from 1.3 to 2.7 sec. The SNR remained fixed throughout at  $-3$  dB. The target sentence presented to the left ear was spoken by a female talker, and the target sentence that was simultaneously presented to the right ear was spoken by a male talker. The number of words in the sentences varied, with on average of six words per sentence (Versfeld et al. 2000).

After the offset of the masking noise, participants were tasked to (1) attend to the sentence spoken by the female (left) and ignore the male talker (right), referred to as the single-sentence condition, or (2) to attend to both speakers simultaneously, referred to as the dual-sentence condition. Thus, in the single-sentence conditions they had to repeat one sentence, but in the dual-sentence conditions they had to repeat two sentences, one from the left ear followed by one from the right ear. Information about which ear to attend to was given to participants verbally before the task block, with no additional auditory or visual cues provided during the task. Participants were encouraged to guess if they were not confident about their answers. Scoring was conducted live during the experiment by an experimenter who was seated outside the sound-treated room. The scoring was based upon all words in the target sentences. Errors in word repetition were not permitted (for instance incorrect tenses, articles, or singular/plural errors), however, errors in word order were treated more leniently (correct words repeated in the wrong order were scored as correct). The scoring was recorded as a proportion of words correctly repeated, out of the total target words in the sentence. There was one target sentence per trial in the single-sentence condition and two target sentences per trial in the dual-sentence condition. There were therefore approximately twice as many words to recall as in the dual-sentence condition as compared with the single-sentence condition.

**VR Manipulation: Agents and Nonagent Controls** • The virtual scene was manipulated by the introduction of agents or nonagent controls. Within the headset, the virtual environment

\*The two repeated condition orders were as follows: (1) dual sentence/nonagent controls, (2) single sentence/agents, (3) dual sentence/agents, (4) single-sentence/nonagent controls; and (1) dual sentence/nonagent controls, (2) single-sentence/nonagent controls, (3) dual sentence/agents, (4) single-sentence/agents.

was 9 meters by 6 meters wide and consistent with the physical laboratory environment, contained a centrally placed chair and a table. Unlike the physical laboratory, the virtual scene contained two additional chairs, positioned 4.5 meters away from the participant's chair, and each 1 meter to the left or right. There were also two windows (out of the view of participants when gazing forward) and a permanent red dot on the back wall, upon which participants were told to direct their gaze during sentence presentation. The participant had no body in the virtual environment. Participants undertook the task in two conditions: (1) in the presence of two agents who gave occasional performance feedback or (2) in the presence of two nonagent controls. Figure 1 shows the virtual environment in both of these conditions.

The agents' physical features were downloaded from Mixamo (<https://www.mixamo.com>, Mixamo characters selected: David and Megan). The characters, one male and one female, were selected based upon their level of human resemblance, in addition to their age demographic and casual clothing which could reasonably place them as peers of our participants. Both agents were seated on chairs in the virtual scene, within the participant's eyeline. They displayed looped, somewhat natural-looking idle behavior, such as breathing and small movements, rather than being completely still. When present, the agents occasionally provided performance feedback to the participant. Feedback lasted for 2 sec in duration.

Positive feedback consisted of an animation of the agent/s nodding their head, and negative feedback consisted of an animation of the agent/s shaking their heads. In a trial, participants could receive feedback from either the male agent, the female agent, or both together. Opposing feedback was never given by the agents (i.e., simultaneously positive and negative feedback). To receive positive feedback, participants had to successfully repeat a full sentence during the single-sentence condition, or two words of each sentence during the dual-sentence condition. If these criteria were not met, negative feedback could be given.

The frequency of feedback was matched approximately equally between the two dichotic conditions, as well as between participants. In general, the chance of feedback being provided by one of the agents in a trial was 15%. However, to account for the increased likelihood of participants earning positive feedback in the single-sentence condition, the chance of negative feedback being given was increased to 80%, relative to positive feedback during that condition. To ensure a similar amount of feedback between conditions, the agents were each programmed to show a maximum of three negative and five positive feedback animations per block. As expected, there was a significant effect of feedback type (positive versus negative) on the number of feedback responses given (i.e., the sum of both agents' responses), such that there were more positive responses given than negative [ $F(1,25) = 115.54, p \leq 0.001$ ,

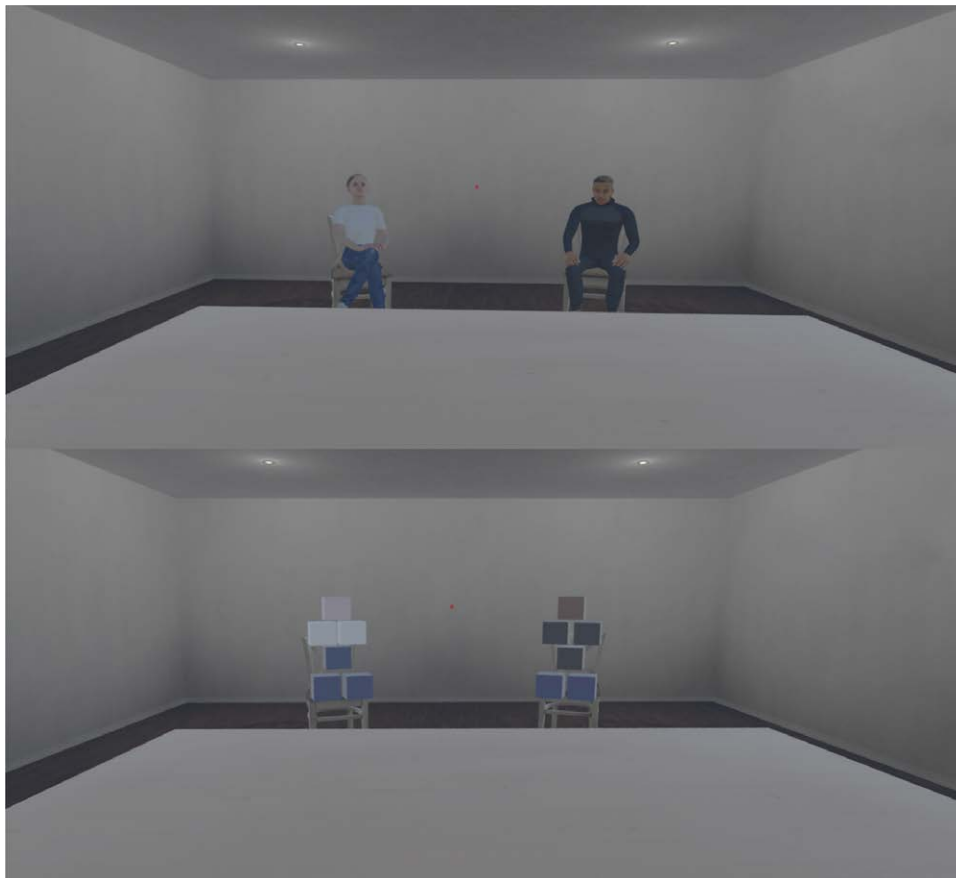


Fig. 1. Screenshots of the virtual reality environment visible to the participant. The upper panel demonstrates the two agents who provided occasional performance feedback, whereas the lower panel demonstrates the nonagent controls who gave no feedback. All other aspects of the room were identical in both the agent and nonagent control conditions.

$\eta_p^2 = 0.82$ ]. In addition, an interaction effect was present [ $F(1,25) = 23.07, p \leq 0.001, \eta_p^2 = 0.48$ ], suggesting that the difference between positive and negative feedback was more pronounced in the single-sentence, rather than dual-sentence condition. This is likely to result from the higher performance in this condition than in the dual-sentence condition—there were fewer opportunities for negative feedback to be given. There was no significant difference between the amount of negative and positive feedback given by the male compared with the female agent [in dual-sentence condition,  $F(1,25) = 2.44, p = 0.13, \eta_p^2 = 0.09$ ; in single-sentence condition,  $F(1,25) = 0.72, p = 0.40, \eta_p^2 = 0.03$ ].

The nonagent controls, on the other hand, consisted of boxes matching a similar color scheme to the agents. They also displayed idle behavior (moving a little, to match the agent idle movements) but did not provide any performance feedback. These nonagent controls were introduced such that we could differentiate whether any effects were resulting from the presence of an object in the visual field, versus an effect of the agent appearance plus feedback.

### Baseline Videos

Three-minute baselines were implemented before each task block, such that reactivity (delta) scores could be calculated for the cardiovascular measures. The purpose of these periods was to allow participants to enter a resting state (Jennings et al. 1992). During these baseline periods, participants watched videos projected onto part of the back wall of the virtual environment. The agents and nonagent controls were present during the baselines preceding the tasks blocks. The videos consisted of footage that was shown for the baselines of a previous experiment (Plain et al. 2021). The videos lasted for 3 min and depicted drone footage shot over Edinburgh's countryside and cityscape. They were neutral in nature, containing no emotive, distressing, or stressful content.

### Subjective Ratings and Questionnaires

After each block, the head-mounted display was removed, and participants completed four separate rating scales about the task using pen and paper. They were asked to rate the following, "How much effort did it take you (on average) to understand the speech?," "How difficult did you find the task?," "Did you understand very few or almost all sentences?," and "To what extent did you feel included in (engaged with) the experiment?." To answer, participants selected their response on a paper visual analog scale that ranged from 1 to 10, with one decimal precision. Each scale had five written anchors, of which the extremes were labeled as follows: "no effort" to "very much effort," "not at all difficult" to "very difficult," "no sentences understood" to "all sentences understood," and finally, "not included at all" to "fully included." After that, the head-mounted display was adjusted again.

After the experiment, participants also completed two questionnaires on paper: the i-group presence questionnaire (Schubert et al. 2001) and a custom-made questionnaire, referred to here as the agent questionnaire. The i-group presence questionnaire measured the user's subjective sense of presence during the experiment. Fourteen items were presented, each consisting of a statement about general presence, spatial presence, involvement, and experienced realism. For example, "I was completely

captivated by the virtual world." The response format consisted of a five-point Likert scale with labeled anchors at the extremes only (e.g., fully disagree to fully agree).

The agent questionnaire aimed to provide more information about the perception of the agents themselves. Five items were presented on a seven-point Likert scale with the labeled extremes presented in brackets here: (1) I was aware of the presence of the agents (strongly agree to strongly disagree); (2) it was like the agents were real (strongly agree to strongly disagree); (3) because of the presence of the agents I felt (less to more) relaxed; (4) because of the presence of the agents I felt (encouraged to discouraged); and (5) compared with when the agents were absent, when they were there I tried (less to more). Additional answer space was provided so that participants could elaborate on their answers if necessary.

### Cardiovascular Data Collection

Cardiovascular data were collected throughout the experiment by a Cardioscreen 2000 system (Medis, Ilmenau, Germany). The system measured ECG and ICG, both at a sampling frequency of 1000 Hz. These signals were obtained by three disposable electrodes positioned on the participant's neck and chest. More specifically, a dual sensor was placed on the left side of the neck, and two single sensors were placed on the left side of the chest: one at the level of the xiphoid process and one 10 cm beneath. Participants wore a blood pressure cuff on their right arm, placed over the brachial artery. The blood pressure cuff was inflated once per baseline and once per task block (at around the 24th sentence), with around 6 min between successive inflations. The inflation of the blood pressure cuff lasted on average of 30 sec.

### Cardiovascular Measures

For the purposes of our analysis, we extracted cardiovascular measures using two different methods. In the first approach, referred to here as the block-wise method, the measures were calculated using an entire task block of data, averaged, and compared with the baseline values. Using this method, we extracted HRV, PEP, heart rate, and blood pressure, generating one reactivity score per condition. The included measures and the direction of expected effort-related effects can be seen in Table 1. This is the more "traditional" approach to cardiovascular psychophysiological measures. However, some studies have suggested that there are fluctuations in cardiovascular measures

**TABLE 1. Included cardiovascular measures, abbreviations, and the expected effort-related change**

Measure	Abbreviation	Expected Effort-Related Change
High-frequency heart rate variability	HF-HRV	Decrease
Pre-ejection period	PEP	Decrease
Heart rate	HR	Increase
Systolic blood pressure	SBP	Increase
Diastolic blood pressure	DBP	Increase
Mean arterial pressure	MAP	Increase

occurring within a trial (Francis et al. 2016), which may be missed when averaging across the whole block. Therefore, in the second approach, referred to here as the trial level method, we explored the measures when extracted instead at the level of the trial, averaged across 1-sec epochs. Such a short data window precludes HRV and blood pressure measures, which require a longer window of data. Therefore, using this trial-level method we only extracted changes in heart rate and PEP.

### Block-Wise Analysis •

**Heart Rate Variability.** HF-HRV and heart rate were both extracted from the ECG signal. Full baseline and task data were imported to MATLAB (version R2018b). Any segments containing visible artifacts were removed, such that all included data were free from noise. A peak detection function was used to find all R peaks in the signal. Visual inspection of the R peaks demonstrated that the appropriate threshold had been selected and all peaks had been properly detected. Next, the interbeat intervals were loaded into Kubios HRV Standard 3.3.1 (Tarvainen et al. 2014), with artifact correction set to low (threshold: 0.3). Normalized HF-HRV, which is normalized using the ratio of the power in the frequency bands 0.15 to 0.4 Hz (high) and 0.04 to 0.15 (low), was extracted. This resulted in one HF-HRV value per task block and one per baseline. Delta reactivity scores were calculated for HRV by subtracting the baseline from the task values.

**Pre-Ejection Period.** PEP was extracted from the ECG and ICG signals. First, R peaks were detected in the ECG signal and visual inspection confirmed that detection was appropriate. The ICG signal was differentiated, and a low-pass Butterworth filter (order four, cutoff 50 Hz) applied. All individual cycles of the ECG and ICG signals were inspected, and any cycles containing artifacts were excluded. The remaining, artifact-free cycles were ensemble averaged across segments of 60 sec. The ensemble-averaged data were used to find PEP, which refers to the time interval between the R-onset of the ECG signal and the B-point of the ICG (selected using the procedure described by Sherwood et al. [1990]). This procedure resulted in three PEP values per baseline (which lasted 3 min) and from 8 to 14 PEP values per task block (the task duration was unfixed and depended upon the speed at which the participant completed sentence repetition).

To ensure reliability, scoring of PEP was conducted by two separate scorers. The PEP values from both scorers were compared with one another. Any differences greater than 10 msec were reviewed in a meeting between the scorers, and any errors in scoring were corrected, as necessary. The level of agreement between the two scorers was high: the intraclass correlation coefficient (two-way mixed, absolute agreement) was 0.98 before the scorers' meeting and 0.99 after errors had been corrected. Final PEP values were calculated as an average of the two scorers' PEP scores. PEP was averaged across the full

3-min baseline period and from minutes one to eight of the task period. This time period was selected for the task to ensure that all participants' PEP score contained an average of the same number of minutes, because participants completed the task blocks at different times. The shortest task period was 8 min in duration. Lastly, delta reactivity scores were calculated by subtracting the average baseline PEP from the average task PEP.

**Heart Rate and Blood Pressure.** Heart rate was determined from the interbeat intervals obtained during HF-HRV extraction (Heart rate variability). The interbeat intervals were converted to heart rate (in beats per minute) and mean heart rate was calculated for each condition. Three individual blood pressure measures were extracted per condition: mean arterial pressure (MAP), systolic blood pressure (SBP), and diastolic blood pressure (DBP). Delta reactivity scores for heart rate and blood pressure were calculated by subtracting baseline values from the task values. Trials coinciding with blood pressure cuff inflation were included in the analysis.

**Exploratory Trial Level Analysis •** Two participants were excluded from the exploratory trial level analysis due to an issue with triggering that compromised detection of their data corresponding to individual trials. Therefore, 24 participants were included in the analysis of heart rate and 22 participants were included in the analysis of PEP. Using MATLAB, ECG, and ICG data were filtered using sixth-order Butterworth band-pass filters with a passband of 1 to 30 Hz (Raza et al. 1992). The data corresponding to each trial were extracted from 1 sec before trial onset until 9 sec after trial onset. This 10-sec analysis window spanned from 1 sec before the presence masking noise to the approximate end of the postsentence masking noise (shaded in Fig. 2). Data after this (i.e., during the response time, scoring, and any feedback presentation) were not included in the exploratory cardiovascular analysis.

After the 1 sec of silence, the trials were split into three sections, based upon the approximate acoustic sections of the trial. The first 3 sec always corresponded to the masking noise presentation alone. This section will be referred to as presence masking noise. Sentence presentation varied from 1.3 to 2.7 sec in duration. To ensure all the sentences (and any instantaneous physiological changes elicited by sentence presentation) were adequately captured in our time window, we considered the next 3 sec to correspond to the dichotic sentence presentation. Last, the remaining 3 sec were considered to represent the postsentence masking noise. We acknowledge that these latter two-time windows are approximate and may be contaminated with other trial sections, however splitting the data in this way allowed us to compare timings across trials with different durations.

The data obtained during each trial were split into 1-sec epochs. For each epoch, the R peak of the ECG was detected using the same peak detection function referred to in Heart rate variability. Individual scripts were created to allow customized

Trial structure	Silence	Pre-sentence masking noise	Dichotic sentences presented	Post-sentence masking noise	Participant response	Scoring conducted	Feedback (if any)
Duration (s)	1	3	3*	3	3*	2*	2

Fig. 2. Schematic demonstrating the trial structure. The shaded section (consisting of the silence, presence masking noise, dichotic sentence presentation, and postsentence masking noise) reveals the 10-sec window selected for the exploratory analysis. Trial durations are given in seconds. Those marked with an asterisk are approximate timings, as these sections of the trial had an unfixed duration.



peak detection thresholds, and visual inspection of the signal was carried out to ensure the correct threshold had been chosen. When the correct threshold had been determined, the interbeat interval between this R peak and the subsequent R peak was calculated. If two or more R peaks were present in the epoch, the average of the interbeat intervals was calculated. In this way, one interbeat interval was obtained per one-second epoch. Interbeat intervals were then mathematically converted to heart rate and the values were normalized to the heart rate value calculated during the silence preceding the trial.

The same epoch-based process was conducted for PEP. In contrast with the approach described in Pre-ejection period, trial-level PEP was not obtained by manual scoring. Instead, a semiautomatic procedure was devised and implemented in MATLAB to detect the Q point of the ECG using peak detection, and the B-point of the ICG using a tangential method. Visual inspection was conducted to ensure adequate positioning of these points. The distributions of the heart rate and PEP values obtained from each participant were reviewed, and any outliers that were considered to be physiologically unlikely were excluded. Values were averaged across each condition (30 trials), such that each participant had nine average heart rate and PEP values, representing the proportional change from pretrial value for each second of the trial.

For both heart rate and PEP, the areas under the curve were calculated in MATLAB using trapezoidal numerical integration. Calculating area under the curve can allow statistical analysis of discrete-time course data (Jaki & Wolfsegger 2009). The advantage of doing so here was that the distinct parts of the trial could be analyzed without averaging across time and potentially losing important information. An increase in the area under the curve was interpreted to reflect an increase in heart rate or PEP, respectively.

**Statistical Analysis**

Statistical analysis was conducted in IBM SPSS statistics (Version 28). Two-way repeated measures ANOVAs were conducted to determine effects (main or interaction) of the task demand condition (single- or dual-sentence) and the VR manipulation (agents or nonagent controls) on performance, cardiovascular reactivity measures, self-reported effort, difficulty, performance, and engagement. The area under the curves of the three trial sections (presentence masking, dichotic sentence presentation, and postsentence masking) was also analyzed by repeated measures ANOVAs. To account for the inclusion of multiple cardiovascular measures in the block-wise analysis and three different time points in the trial-level analysis, all cardiovascular *p* values were corrected for multiple comparisons using the false discovery rate Benjamini–Hochberg correction (Benjamini & Hochberg 1995; Bird & Hadzi-Pavlovic 2014; Martínez-Cagigal 2021).

**RESULTS**

**Performance**

The total number of words correctly repeated in each condition is reported in Table 2. For statistical analysis, the proportion of words correctly repeated was calculated by dividing the sum of the correct words by the total number of presented words for each trial (there were twice as many target words for the dual-sentence condition compared with the single-sentence condition). The average proportional performance in each condition is presented in Table 2 and Figure 3. A two-way repeated measures ANOVA was conducted to assess the effects of task demand condition and VR manipulation on proportional performance. The results revealed a significant effect of task demand [ $F(1,25) = 163.55, p < 0.001, \eta_p^2 = 0.87$ ], such that

**TABLE 2. Average performance, subjective ratings, and cardiovascular data with SEs in brackets**

	Agent		Nonagent Control	
	Single Sentence	Dual Sentence	Single Sentence	Dual Sentence
<b>Performance</b>				
Proportion of words correct	0.95 (0.01)	0.68 (0.03)	0.95 (0.01)	0.67 (0.02)
Total words correct	5.91 (0.05)	8.33 (0.30)	5.84 (0.09)	8.30 (0.31)
<b>Subjective rating</b>				
Effort	4.41 (0.30)	7.78 (0.23)	4.26 (0.30)	7.74 (0.28)
Difficulty	3.36 (0.28)	7.50 (0.27)	3.18 (0.28)	7.57 (0.26)
Performance	8.55 (0.19)	5.55 (0.25)	8.64 (0.15)	5.37 (0.29)
Engagement	6.95 (0.56)	7.15 (0.41)	6.54 (0.55)	6.84 (0.46)
<b>Cardiovascular baseline data</b>				
HF-HRV (n.u.)	44.72 (2.99)	45.06 (3.85)	49.30 (3.90)	45.94 (3.38)
PEP (msec)	103.32 (2.26)	103.60 (2.17)	103.12 (2.26)	103.72 (2.22)
Heart rate (bpm)	69.75 (1.68)	70.43 (1.90)	68.55 (1.81)	69.32 (1.81)
SBP (mm Hg)	118.31 (2.19)	118.54 (1.99)	117.96 (1.88)	118.54 (2.03)
DBP (mm Hg)	73.19 (1.30)	73.88 (1.25)	73.46 (1.26)	72.81 (1.16)
MAP (mm Hg)	82.92 (1.39)	83.12 (1.34)	83.58 (1.28)	83.08 (1.27)
<b>Cardiovascular reactivity data</b>				
HF-HRV (n.u.)	-8.35 (3.16)	-10.59 (2.68)	-8.67 (3.39)	-17.24 (2.63)
PEP (msec)	0.19 (0.45)	0.30 (0.57)	0.60 (0.53)	0.82 (0.54)
Heart rate (bpm)	2.85 (0.53)	2.55 (0.63)	3.71 (0.42)	5.00 (0.72)
SBP (mm Hg)	4.50 (1.41)	5.23 (0.95)	5.50 (0.82)	6.08 (1.02)
DBP (mm Hg)	3.35 (0.49)	3.19 (0.46)	3.38 (0.54)	4.50 (0.82)
MAP (mm Hg)	3.50 (0.49)	3.81 (0.56)	2.54 (0.39)	4.46 (0.83)

Subjective rating scales ranged from 1 to 10, with one decimal precision. PEP reactivity values are expected to be negative, suggesting more effort investment during the task compared with the baseline, however, the opposite was true here. This should be interpreted cautiously however, as the values are very small (all below 1 msec) and the SE values overlap zero in some cases. DBP, diastolic blood pressure; HF-HRV, high frequency heart rate variability; MAP, mean arterial pressure; PEP, pre-ejection period; SBP, systolic blood pressure.

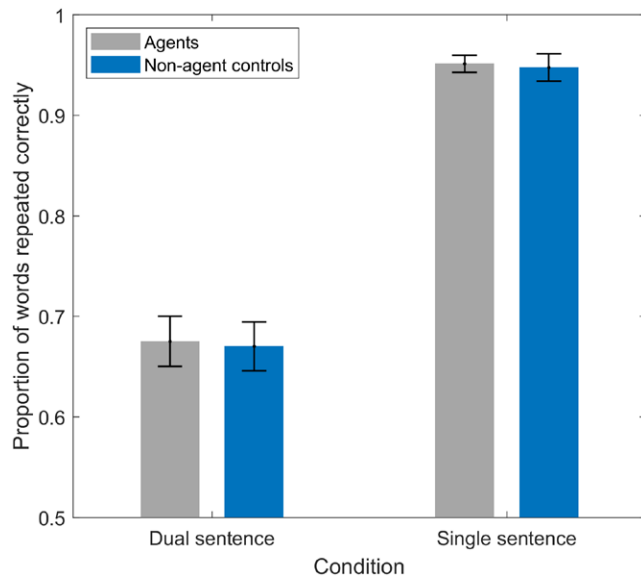


Fig. 3. The average proportion of words repeated correctly in the dual- and single-sentence conditions, in the presence of the agents and nonagent controls. Error bars represent SE of the mean.

proportional performance was higher in the single-sentence condition compared with the dual-sentence condition. There was no main effect of the VR manipulation [ $F(1,25) = 0.22$ ,  $p = 0.64$ ,  $\eta_p^2 = 0.01$ ] and no interaction between task demand and the VR manipulation [ $F(1,25) = 0.01$ ,  $p = 0.93$ ,  $\eta_p^2 = 0.00$ ]. A scatter plot demonstrating the summed proportion of words repeated correctly from each speaker in the dual-sentence condition indicated that the participants preferentially repeated words presented to the left ear as compared with the right ear (Fig. 4). This observation is consistent with previous studies that implemented a similar design (Best et al. 2010; Koelewijn et al. 2014).

### Subjective Ratings and Questionnaires

After each block, participants rated their perceived effort investment, the difficulty of the task, their performance, and engagement in the task. The means and SEs for each rating scale are presented in Table 2. Repeated measures ANOVAs (results displayed in Table 3) demonstrated that subjective effort investment and the difficulty of the task increased, while subjective performance decreased, in the dual-sentence compared with single-sentence condition. The VR manipulation did not have a significant effect on any of the subjective ratings. In addition, there were no significant effects of the task demand condition, VR manipulation, nor any interaction between the two on the participants' reported level of engagement.

When all blocks of the dichotic listening task had been completed, participants completed the two questionnaires. The results of the i-group presence questionnaire were summed for each participant. The mean score across participants was 41.36 (SD = 5.45). Boxplots of the agent questionnaire results are demonstrated in Figure 5.

### Block-Wise Cardiovascular Analyses

All 26 participants were included in the cardiovascular analyses, except for two with unreliable ICG signals, precluding

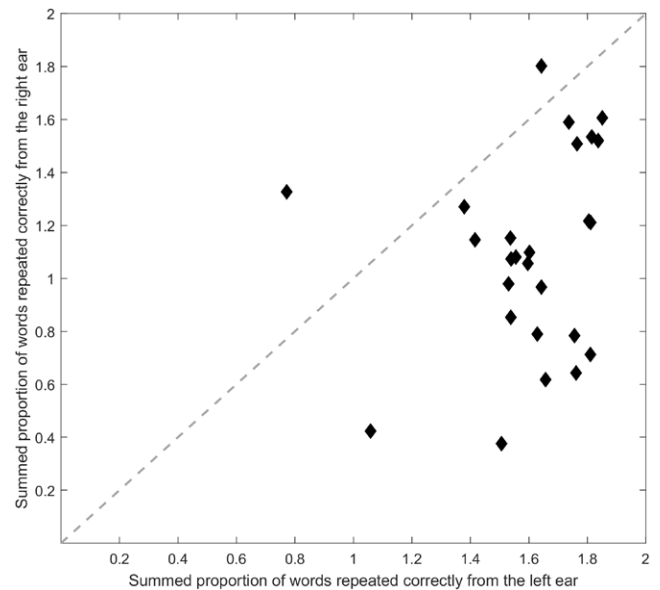


Fig. 4. A scatter plot of the summed average proportion of words repeated correctly from the left ear compared with the right ear. The proportion of words in the agent and nonagent control conditions are added together to make a single score per participant. Each diamond represents one participant.

them from inclusion in the PEP analysis. Average cardiovascular baseline values and SEs for each measure are presented in Table 2. In studies of cardiovascular reactivity, the baseline data are often analyzed to determine any effects of the experimental manipulations on the baseline periods (Richter et al. 2008; Richter & Gendolla 2009). This was deemed particularly necessary here as the agents and nonagent controls were present during the baselines. Despite this, no significant effects were demonstrated in the baseline data (all corrected  $p$  values  $>0.05$ ).

Cardiovascular reactivity data are presented in Table 2 and Figure 6. Pearson correlation coefficients revealed that no cardiovascular variables were correlated significantly with BMI, therefore BMI was not included or accounted for in the analysis. Repeated measures ANOVAs were conducted on the reactivity scores to determine the presence of any effects of the VR manipulation or task demand condition on HF-HRV, PEP, heart rate, SBP, DBP, and MAP. Results of this analysis are presented in Table 4. A significant effect of the VR manipulation was demonstrated on heart rate, such that heart rate was higher in the nonagent control conditions compared with the agent conditions. Otherwise, no significant effects were demonstrated on any cardiovascular variables.

### Exploratory Trial Level Cardiovascular Analyses

Figure 7 reveals the pattern of average heart rate and PEP over the trials in each of the different conditions. The area under the curves of heart rate and PEP in the three different time windows (presence masking, dichotic sentence presentation, and postsentence masking) are presented in Table 5. Repeated measures ANOVAs were conducted to demonstrate any effects of the VR manipulation or dichotic listening on these measures within each time interval. Results of this analysis are demonstrated in Table 6. A significant effect of the task demand condition on heart rate was demonstrated in

**TABLE 3. Results of repeated measures ANOVAs for subjective rating data**

	Effect	$F(1,24)$	$p$	$\eta_p^2$
Effort	Task demand	<b>238.53</b>	<b>&lt;0.001</b>	<b>0.91</b>
	VR manipulation	0.28	0.60	0.01
	Interaction	0.10	0.76	0.00
Difficulty	Task demand	<b>214.52</b>	<b>&lt;0.001</b>	<b>0.90</b>
	VR manipulation	0.16	0.69	0.00
	Interaction	0.66	0.43	0.03
Performance	Task demand	<b>310.02</b>	<b>&lt;0.001</b>	<b>0.93</b>
	VR manipulation	0.08	0.78	0.00
	Interaction	0.57	0.46	0.02
Engagement	Task demand	0.79	0.38	0.03
	VR manipulation	4.02	0.06	0.14
	Interaction	0.14	0.71	0.00

Significant effects are denoted by bold font.  
ANOVA, analysis of variance.

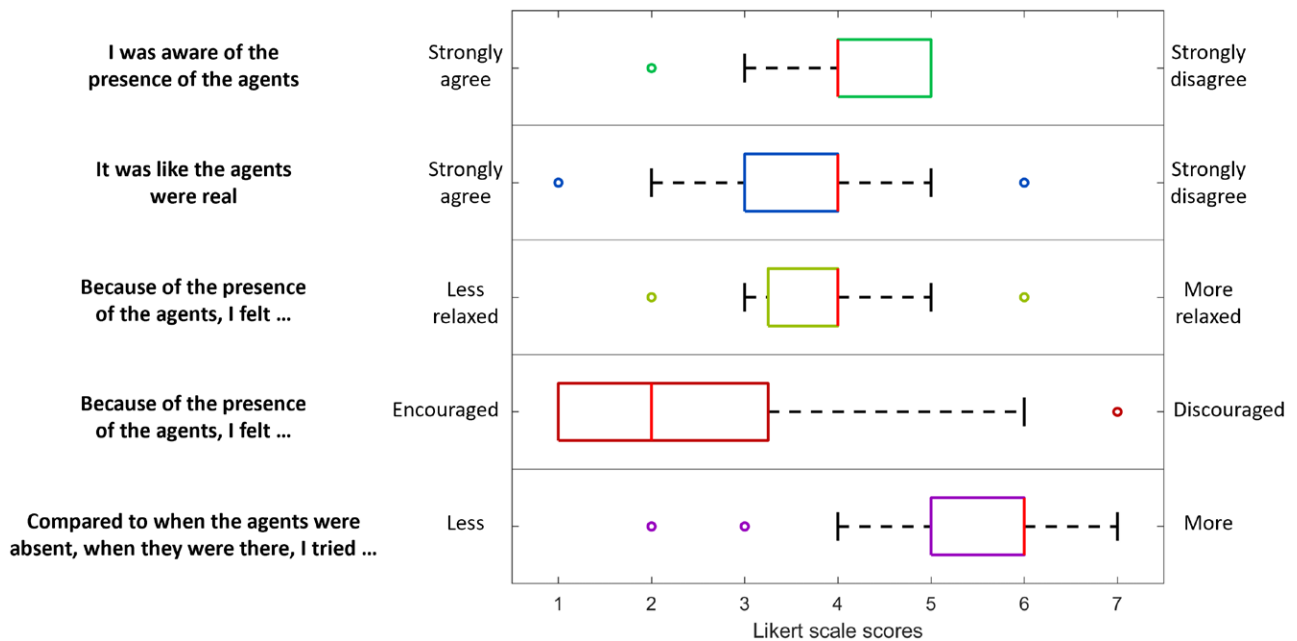


Fig. 5. Boxplot demonstrating the spread of answers for the agent questionnaire.

the presentence masking noise [ $F(1,23) = 12.16, p = 0.02, \eta_p^2 = 0.35$ ] and dichotic sentence presentation phases [ $F(1,23) = 8.62, p = 0.03, \eta_p^2 = 0.27$ ]. This reveals that in these two-time windows, there was a greater area under the curve (i.e., heart rate was higher) in the dual-sentence condition, compared with the single-sentence condition. Otherwise, no significant effects were demonstrated in any time window, on heart rate or PEP.

**DISCUSSION**

The present study investigated the impact of incorporating VR agents in a dichotic speech reception task. Outcome measures included task performance, subjective ratings (effort, task difficulty, performance, and engagement), and cardiovascular measures. We used two different methods of cardiovascular analysis: a block-wise analysis, where HF-HRV, PEP, heart rate, and blood pressure were computed across complete task blocks, and an exploratory trial-level investigation into heart rate and PEP. We hypothesized that more effort would be invested in

the dual compared with the single-sentence condition, causing increased cardiovascular reactivity. In addition, we expected the presence of and evaluation by the agents to increase success importance, and therefore increase the cardiovascular response, compared with the nonagent control conditions.

**VR Agents and Nonagent Controls**

The main finding of the block-wise cardiovascular analysis was an effect of the VR manipulation on heart rate. However, the direction of this effect contradicted our expectations. Heart rate was higher in the nonagent control conditions compared with the conditions with agents present. This finding occurred in isolation: it was not accompanied by similar changes in performance, subjective ratings, or any other cardiovascular measures (HF-HRV, PEP, or blood pressure measures).

There are two primary aspects to discuss regarding these cardiovascular findings. The first is the lack of evidence for the expected effect (i.e., why was there no measurable decrease in HF-HRV or increase in PEP, heart rate, SBP, DBP, and MAP

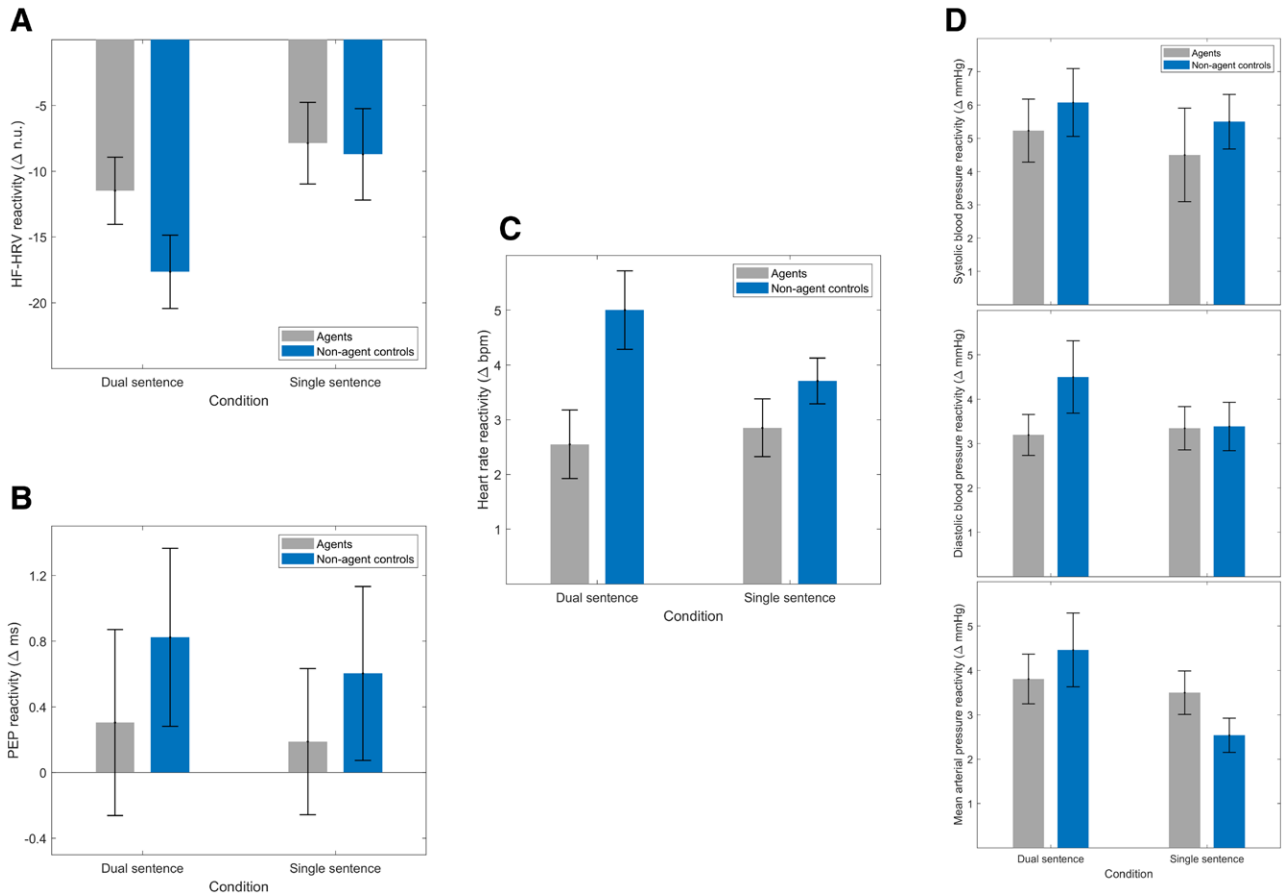


Fig. 6. Average cardiovascular reactivity scores with SE bars. (A) HF-HRV, (B) PEP, (C) heart rate, and (D) blood pressure: systolic blood pressure (upper), diastolic blood pressure (middle) and mean arterial pressure (lower). bpm indicates beats per minute; HF-HRV, high-frequency heart rate variability; mm Hg, millimeters of mercury; msec, milliseconds; n.u., normalized units; PEP, pre-ejection period.

TABLE 4. Results of repeated measures ANOVAs for cardiovascular measures

	Effect	<i>F</i> (1,25)	<i>p</i>	$\eta_p^2$
HF-HRV	Task demand	3.83	0.36	0.13
	VR manipulation	1.84	0.41	0.07
	Interaction	2.08	0.41	0.08
PEP	Task demand	0.23	0.77	0.01
	VR manipulation	3.29	0.36	0.13
	Interaction	0.01	0.94	0.00
Heart rate	Task demand	1.47	0.41	0.06
	VR manipulation	<b>25.04</b>	<b>&lt;0.001</b>	<b>0.50</b>
	Interaction	2.70	0.40	0.10
SBP	Task demand	0.35	0.72	0.01
	VR manipulation	1.41	0.41	0.05
	Interaction	0.01	0.94	0.00
DBP	Task demand	1.12	0.45	0.04
	VR manipulation	1.38	0.41	0.05
	Interaction	0.95	0.47	0.04
MAP	Task demand	3.96	0.36	0.14
	VR manipulation	0.08	0.88	0.00
	Interaction	1.91	0.41	0.07

Significant effects are denoted by bold font. Corrected *p* values are presented. Please note degrees of freedom for PEP analysis were (1,23) due to missing data from two participants. DBP, diastolic blood pressure; HF-HRV, high frequency heart rate variability; MAP, mean arterial pressure; PEP, pre-ejection period; SBP, systolic blood pressure.

in the presence of the agents?) and the second is the presence of the opposite effect (i.e., why did heart rate increase more in the absence of the agents?). A potential reason for the former may be due to the realism of the agents. Around half (46%)

of the 24 participants who provided additional comments in the free written part of the agent questionnaire explicitly commented in the free written part of the agent questionnaire that the agents were not realistic in their movements, interaction, or



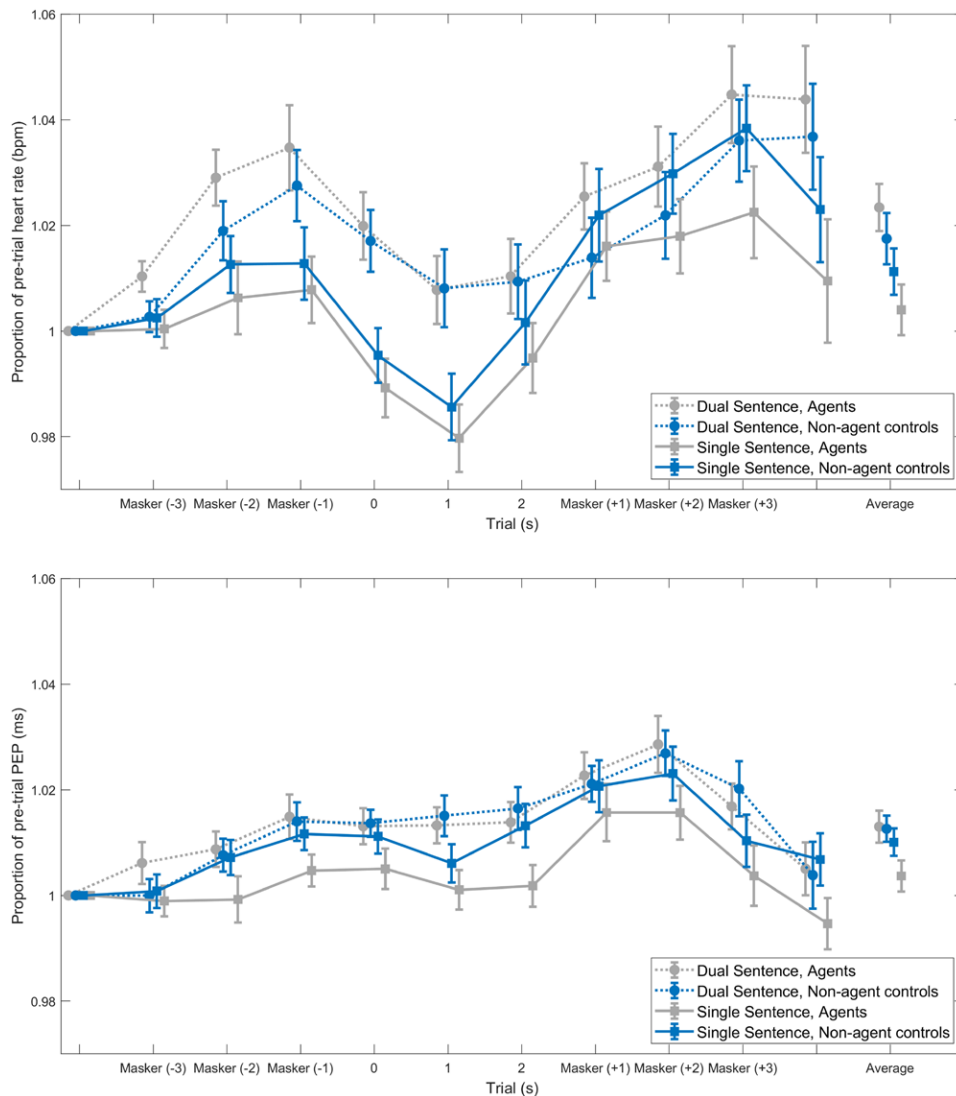


Fig. 7. Average heart rate (upper panel) and PEP (lower panel) in 1-sec intervals, presented as a proportion of the value recorded in the silence preceding the onset of the presence masking noise. Error bars reflect the SE of the mean. bpm indicates beats per minute; msec, milliseconds; PEP, pre-ejection period.

TABLE 5. Area under the curve averages and SEs

	Agent		Nonagent Control	
	Single Sentence	Dual Sentence	Single Sentence	Dual Sentence
<b>Heart rate</b>				
Masking only	6.03 (0.04)	6.16 (0.03)	6.07 (0.03)	6.11 (0.03)
Sentence presentation	5.91 (0.04)	6.07 (0.04)	5.96 (0.04)	6.07 (0.04)
Postsentence masking	6.11 (0.04)	6.19 (0.04)	6.19 (0.04)	6.14 (0.05)
<b>PEP</b>				
Masking only	6.01 (0.02)	6.05 (0.02)	6.03 (0.02)	6.04 (0.02)
Sentence presentation	6.03 (0.02)	6.08 (0.02)	6.05 (0.03)	6.09 (0.02)
Postsentence masking	6.09 (0.03)	6.14 (0.03)	6.10 (0.03)	6.14 (0.02)

PEP, pre-ejection period.

facial expressions. Some examples of comments on the agent questionnaire item 2 (“It was like the agents were real”), translated from Dutch to English include: “they did not resemble real humans and I knew it was a VR world,” “I could clearly notice that they were not real,” “they were moving too robotically” and “mainly the faces were unrealistic.”

A lack of realism may have detracted from the participants’ feeling of presence in the virtual environment (Baños et al. 2005; Diemer et al. 2015). This is supported by the i-group presence questionnaire results, which were lower than have been demonstrated in other studies (Vasconcelos-Raposo et al. 2016; Hruba et al. 2020). To increase presence, studies using

**TABLE 6. Area under the curve repeated measures ANOVA analysis**

Measure	Time Window	Effect	$F(1,23)$	$p$	$\eta_p^2$
HR	Presentence masking	Task demand	<b>12.16</b>	<b>0.02</b>	<b>0.35</b>
		VR manipulation	0.13	0.72	0.00
		Interaction	3.24	0.19	0.12
	Sentence presentation	Task demand	<b>8.62</b>	<b>0.03</b>	<b>0.27</b>
		VR manipulation	1.05	0.57	0.04
		Interaction	0.42	0.71	0.02
	Postsentence masking	Task demand	0.24	0.71	0.01
		VR manipulation	0.29	0.71	0.01
		Interaction	3.79	0.19	0.14
Measure	Time Window	Effect	$F(1,21)$	$p$	$\eta_p^2$
PEP	Presentence masking	Task demand	0.94	0.52	0.04
		VR manipulation	0.01	0.94	0.00
		Interaction	1.21	0.52	0.05
	Sentence presentation	Task demand	4.30	0.27	0.17
		VR manipulation	1.47	0.52	0.07
		Interaction	0.02	0.94	0.00
	Postsentence masking	Task demand	4.00	0.27	0.16
		VR manipulation	0.98	0.52	0.05
		Interaction	0.05	0.94	0.00

Significant effects are denoted by bold font. Corrected  $p$  values are presented. ANOVA, analysis of variance; HR, heart rate; PEP, pre-ejection period.

VR versions of the TSST have included evaluators who speak (Shiban et al. 2016), write notes on the paper in front of them (Standard et al. 2020), move their gaze and reposition their feet (Liu & Zhang 2020). One study even incorporated real-time eye tracking of the participant, to allow eye contact between the participant and virtual audience members (Zimmer et al. 2019). More realistic animations in the present study may have been more convincingly elicited a sense of “presence” for the participants.

Another potential reason for a lack of evidence for the expected effects in the cardiovascular measures relates to the timings of the stimuli presented in this study. The proportion of each trial dedicated to listening to the presented sentence was small: just 1.84 sec on average, compared with 6 sec of masking noise (3 sec before and after the dichotic sentence presentation), the participant response time, scoring time and finally, 2 sec of agent feedback. The additional nonsentence presentation time during each trial may have added additional noise to the cardiovascular signals, obscuring any true response. Supporting this theory, a previous study demonstrated that PEP was more sensitive to changes in task demand when extracted during the dichotic sentence presentation of a speech-in-noise trial, rather than averaged across whole task blocks (Plain et al. 2020).

The second point for discussion is that the heart rate response was larger in the nonagent control conditions compared with the agent conditions, or phrased in the opposite way, the heart rate response was smaller with the agents present. This finding is surprising as generally social presence (physical and virtual) is associated with either increased or unchanged heart rate (Wright et al. 1998; Gendolla & Richter 2006; Fairclough & Roberts 2011; Plain et al. 2021). For example, in studies incorporating a physical social presence, heart rate increased while participants were evaluated (Wright et al. 1998) or given performance feedback (Fairclough & Roberts 2011), yet remained unaffected by observation (Gendolla & Richter 2006; Plain et al. 2021). Similarly, some studies incorporating

virtual social agents in the TSST revealed increases in heart rate (Shiban et al. 2016; Zimmer et al. 2018; Fallon et al. 2021), though the response was smaller than that elicited by the in person TSST.

Here, a smaller response with the agents present could be attributed to a supportive or reassuring effect of the evaluation or feedback provided by the agents. In the nonagent control condition, participants received no performance feedback, whereas in the agent condition, they were shown occasional positive or negative animations reflecting whether they were correct or incorrect. Positive feedback was given more often than negative feedback: up to five positive animations and three negative animations were shown per block of 30 sentences. The feedback may have provided reassurance to the participant that they were succeeding at the task. Indeed, the results of the agent questionnaire (Fig. 5) suggest that generally the presence of the agents was encouraging, though there was a significant spread across the Likert Scale. Some studies suggest that social support reduces heart rate (Vrana & Rollock 1998; Ditzen et al. 2007), though it should be noted that these studies included physical touch, which is not directly relevant here.

Another option is that the participants may have simply been more distracted by the animations of the agents with feedback than the nonagent controls (i.e., the measured effect may have been unrelated to social-evaluative threat and purely due to a calming visual distraction). Indeed, during some medical or dental procedures, visual distraction in VR is used as a tool to reduce anxiety levels (Van Twillert et al. 2007; Kaur et al. 2021; Bernaerts et al. 2022). It is interesting that there was no measurable difference in performance between the agents and nonagent control conditions, suggesting that participants were not distracted sufficiently to affect their performance level.

Alternatively, a greater physiological response to the nonagent controls may have resulted from their somewhat uncanny resemblance to humans, which may have triggered evaluative threat. Future work would benefit from the separate manipulation of visual distraction, VR agents, and feedback, to allow

more definitive conclusions to be made about the origins of any physiological changes seen. Future work may also benefit from the inclusion of more in-depth subjective rating items, providing a deeper understanding of participants' perception of the agents. For example, other authors have included items specifically related to copresence, embarrassment, and likeability (Bailenson et al. 2005).

### Dichotic Listening Conditions

In keeping with the results of other dichotic speech reception studies, performance was poorer in the dual-sentence condition, compared with the single-sentence condition (Best et al. 2010; Koelewijn et al. 2018). In the present study, stimuli were presented at  $-3$  dB, matching one of the conditions applied by Koelewijn et al. (2014). At this SNR, Koelewijn et al.'s participants obtained average performance scores of 62% for the dual-sentence condition and 89% for the single-sentence condition, not dissimilar from the performance measured here (around 68% and 95%, respectively). Subjective ratings were also sensitive to the dichotic listening condition: participants recognized that their performance decreased, while their subjective effort investment and perceived task difficulty increased, in the dual-sentence condition compared with the single-sentence condition.

Surprisingly, none of the block-wise cardiovascular measures were affected by the task demand manipulation. In particular, we expected to see a larger decrease in HF-HRV (revealing more effort) in the dual-sentence condition compared with the single-sentence condition, as this measure has been shown to be sensitive to task complexity during a dichotic digit listening task, which also involved dividing attention between the two ears (Seeman & Sims 2015). Notably, Seeman and Sims (2015) measured changes to HRV in response to their dichotic manipulation, even when performance was near ceiling (above 97%) for all conditions. As described earlier, performance in the present study was lower in all conditions, suggesting that the task required effort investment. It could also be argued that the task was more difficult in the present study, because participants were required to repeat whole sentences, rather than digits. In addition to different tasks, that is, digits versus sentence presentation, the dissimilar cardiovascular results between Seeman and Sims' and the present study could be attributed to differences in measures. We calculated the high-frequency component of the HRV (HF-HRV), which is thought to reflect parasympathetic nervous system activity, whereas Seeman and Sims presented the SD of normal-to-normal intervals, which is a more mixed autonomic measure (Malik et al. 1996). It is interesting that Mackersie and Cones' (2011) participants underwent a similar dichotic digit task, with near-ceiling level performance in their different conditions, but no changes in heart rate were revealed.

### Trial Level Analysis

**Heart Rate** • One of the additional aims of this study was to explore patterns of heart rate and PEP within a trial. To achieve this, we conducted an analysis similar to that of Francis et al. (2016), which involved averaging the measures (here, heart rate, and PEP) in 1-sec interval during the trial and normalizing them to pretrial values. The timings of stimuli presented by Francis

et al. differed from those in our study. Francis et al.'s trials began with 0.5 sec of silence, followed by the target sentence (2.8 to 3.5 sec), then 8 sec of silence, before the participants responded. In our study, 3 sec of masking noise preceded and followed the target sentences, and the participant responded immediately after this, without any enforced periods of silence. Then, after the participant's verbal response, the experimenter scored the response, and there was a 2-sec interval for any performance feedback to be given by the agents.

Despite these timing differences, there are some similarities between the patterns of heart rate in the studies. For example, both studies reveal a slight increase in heart rate after the onset of the trial (after sentence onset for Francis et al. [2016], and the masking onset in our study), followed by a subsequent decrease in heart rate. In Francis et al.'s (2016) results, the dip in heart rate occurs at around 7 sec after the sentence onset, whereas this occurs sooner in our study, around 5 sec after masking onset or during the dichotic sentence presentation. A decrease (i.e., slowing) in heart rate is reported in the phasic heart rate literature to reflect anticipation of the response (Jennings & Van Der Molen 2002). We suggest that in our data, heart rate slows as the participants are preparing for the sentence presentation to begin.

To test our original hypotheses on the trial level heart rate and PEP data, we conducted repeated measures ANOVAs on the areas under the curves in three different time windows: the presentence masking, dichotic sentence presentation, and the post-sentence masking. This analysis revealed that during the first two of the three-time windows, there was a significant effect of the task demand on heart rate, demonstrating that the area under the curve was greater (i.e., heart rate increased) in the dual-sentence condition compared with the single-sentence condition. This is of particular interest because acoustically the dual- and single-sentence conditions were identical, which means that the changes demonstrated are not simply due to the level of the sound, as has been shown in other studies (Shoushtarian et al. 2019). The difference between heart rate in the dual- and single-sentence trials in these windows is likely to be caused by the knowledge and anticipation that the dual-sentence trials will be more difficult, compared with the single-sentence condition.

The higher heart rate during the dual-sentence condition compared with the single-sentence condition is then maintained during the presentation of the stimulus but disappears during the latter part of the trial (postsentence masking). This may reflect that the data selected for this part of the analysis was contaminated with other parts of the trial, or even the participant response time. For example, the windows for analysis were always split into 3-sec intervals (3-sec presentence masking, 3-sec dichotic sentence presentation, and 3-sec postsentence masking), regardless of the sentence duration. If sentence presentation was brief in one trial, the data selected for analysis as the postsentence masking noise, may actually reflect mainly the participant response. The lack of a fixed trial duration and an interval of silence before the participant response, are limitations of the present analysis.

The area under the curve heart rate analysis revealed no significant effect of the VR manipulation. This is likely to relate to the timings of the trial, specifically with regards to the feedback timing. Several studies have revealed that providing performance feedback results in heart rate changes that occur almost instantly, affecting two to three interbeat intervals after the

feedback is presented (Somsen et al. 2000; Crone et al. 2003; Van Der Veen et al. 2004). In the present study, the trial level heart rate data were limited to the period before the feedback was given (i.e., the presentation of the presence and post-sentence masking noise and the dichotic sentence presentation). The timing of the feedback presentation is not precisely known, due to unfixed response and scoring durations. It is therefore possible that the feedback caused a short-lived response that was not captured during the trial-level analysis window. Future work would benefit from more optimized, fixed trial structure to capture responses to the feedback. In addition, it may prove informative in future work to explore the responses to different types of feedback (positive, negative, or no feedback trials). Negative feedback, for example, has been shown to cause a slowing of heart rate (Somsen et al. 2000; Jennings & Van Der Molen 2002; Van Der Veen et al. 2004).

**Pre-Ejection Period** • In contrast with heart rate, no significant trial level fluctuations in PEP were elicited by any of the task manipulations, in any time window. This could relate to the timings of stimulus presentation (the relatively short period of sentence presentation in each trial, as discussed earlier) but also the autonomic nervous system origins of PEP. Changes in PEP result from sympathetic nervous system activity on the heart, the influence of which is known to act slower than the parasympathetic nervous system (Draghici & Taylor 2016; Christensen et al. 2020). It is possible that any changes elicited in PEP at the trial level appeared after the time window selected for analysis or impacted the subsequent sentence (and thus the baseline correction of the next sentence).

An additional contributory factor, and a limitation of the present study, may relate to the semiautomated procedure used in the PEP analysis. We devised our own novel algorithm to detect the relevant features of the ECG (Q-point) and ICG (B-point) signals. Though care was taken to visually inspect the output of our algorithm, it is well known that the B-point of the ICG especially can be very difficult to determine accurately in an automated fashion (Sherwood et al. 1990). This is due to signal artifacts caused by body movements and respiration, as well as inherent variability in the morphology of the signal between individuals (Forouzanfar et al. 2018). Several automated procedures for determining PEP have been suggested (Szilágyi et al. 1992; Berntson et al. 2004; Lozano et al. 2007; Arzeno et al. 2008; Árbol et al. 2017), but they do not perform optimally, and therefore expert visual inspection performed by multiple scorers (as was conducted for the block-wise analysis in this study) is the recommended approach (Sherwood et al. 1990; van Lien et al. 2013). The semiautomated tangential method used in this study to detect the B-point differs from other approaches applied in the literature (see examples of some B-point detection algorithms here: Forouzanfar et al. 2018). It was beyond the scope of this work to validate the algorithm itself. Thus, the results of the trial-level PEP analysis should be interpreted with caution.

An additional limitation affecting the trial level analysis is that the precise onset and duration of each blood pressure measurement were not recorded, meaning that trials occurring during blood pressure measurement were included in the analysis. This is a limitation because the blood pressure cuff inflation itself may have had an impact on the participant during the affected trials. For instance, some individuals find blood

pressure cuff inflation uncomfortable (Del-Río-Guerrero et al. 2023). This impact was not anticipated to be substantial, however future trial-level work may benefit from exclusion of blood pressure-contaminated trials.

## CONCLUSIONS

This study measured the cardiovascular reactivity of normally hearing individuals during a dichotic speech reception task in two VR conditions: in the presence of agents who provided performance feedback, or in the presence of nonagent controls, without any performance feedback. Contrary to our expectations, an increase in heart rate was demonstrated during the nonagent control conditions, compared with the agent conditions. This was not reflected in any other cardiovascular measure. In contrast, performance and subjective ratings were sensitive only to the dichotic condition: that is, dual-sentence or single-sentence condition. An exploratory analysis revealed that heart rate fluctuations within a trial differed between the dichotic conditions: heart rate was higher in the first two-thirds of the trial when two sentences had to be repeated, compared with just one sentence. This analysis demonstrates that trial-level cardiovascular measures can be successfully extracted, and analysis of these shorter segments may contain different and supplementary information to that obtained when data are averaged across a block.

## ACKNOWLEDGMENTS

The authors thank J. H. M. van Beek for his support in experimental setup.

This project received funding from the European Union's Horizon 2020 research and innovation program under the Marie-Sklodowska-Curie grant agreement No 765329.

All authors contributed equally to this work. B.P. performed the analysis and wrote the manuscript, H.P. and S.R.B.v.d.V. set up and conducted the experiments, A.A.Z. provided valuable feedback and method conceptualization, M.R. second scored and reviewed the PEP data, T.A.B. assisted with code-writing, S.E.K. applied for funding and is principal investigator of the project. All authors discussed the design, results, and implications and commented on the manuscript at all stages.

This study was approved by the medical ethical research committee of the Amsterdam University Medical Center, location VUmc, reference number: 2018.308.

The authors have no conflicts of interest to disclose.

Address for correspondence: Bethany Plain, Amsterdam UMC, Vrije Universiteit Amsterdam, Otolaryngology Head and Neck Surgery, Ear & Hearing, Amsterdam Public Health Research Institute, Amsterdam, the Netherlands. E-mail: b.plain@amsterdamumc.nl

Received August 30, 2023; accepted August 13, 2024; published online ahead of print October 11, 2024

## REFERENCES

- Ahmed, S. S., Levinson, G. E., Schwartz, C. J., Ettinger, P. O. (1972). Systolic time intervals as measures of the contractile state of the left ventricular myocardium in man. *Circulation*, *46*, 559–571.
- Allen, A. P., Kennedy, P. J., Cryan, J. F., Dinan, T. G., Clarke, G. (2014). Biological and psychological markers of stress in humans: Focus on the Trier Social Stress Test. *Neurosci Biobehav Rev* *38*, 94–124.
- Árbol, J. R., Perakakis, P., Garrido, A., Mata, J. L., Fernández-Santaella, M. C., Vila, J. (2017). Mathematical detection of aortic valve opening (B point) in impedance cardiography: A comparison of three popular algorithms. *Psychophysiology*, *54*, 350–357.
- Arzeno, N. M., Deng, Z.-D., Poon, C.-S. (2008). Analysis of first-derivative based QRS detection algorithms. *IEEE Trans Biomed Eng*, *55*, 478–484.



- Bailenson, J. N., Swinth, K., Hoyt, C., Persky, S., Dimov, A., Blascovich, J. (2005). The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence, 14*, 379–393.
- Baños, R. M., Botella, C., Alcañiz, M., Líaño, V., Guerrero, B., Rey, B. (2005). Immersion and emotion: Their impact on the sense of presence. *Cyberpsychol Behav, 7*, 734–741.
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J Royal Stat Soc Ser B, 57*, 289–300.
- Bernaerts, S., Bonroy, B., Daems, J., Sels, R., Struyf, D., van de Veerdonk, W. (2022). Virtual reality for distraction and relaxation in a pediatric hospital setting: An interventional study with a mixed-methods design. *Front Digit Health, 4*, 133–135.
- Berntson, G. G., Cacioppo, J. T., Quigley, K. S. (1991). Autonomic determinism: The modes of autonomic control, the doctrine of autonomic space, and the laws of autonomic constraint. *Psychol Rev, 98*, 459–487.
- Berntson, G. G., Lozano, D. L., Chen, Y.-J., Cacioppo, J. T. (2004). Where to Q in PEP. *Psychophysiology, 41*, 333–337.
- Best, V., Gallun, F. J., Ihlefeld, A., Shinn-Cunningham, B. G. (2006). The influence of spatial separation on divided listening. *J Acoust Soc Am, 120*, 1506–1516.
- Best, V., Gallun, F. J., Mason, C. R., Kidd, G., Shinn-Cunningham, B. G. (2010). The impact of noise and hearing loss on the processing of simultaneous sentences. *Ear Hear, 31*, 213–220.
- Bird, K. D., & Hadzi-Pavlovic, D. (2014). Controlling the maximum familywise Type I error rate in analyses of multivariate experiments. *Psychol Methods, 19*, 265–280.
- Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annu Rev Psychol, 40*, 109–131.
- Christensen, J., Saunders, G. H., Havtorn, L., Pontoppidan, N. H. (2021). Real-world hearing aid usage patterns and smartphone connectivity. *Front Digit Health, 3*, 111.
- Christensen, J., Saunders, G., Porsbo, M., Pontoppidan, N. (2020). *The everyday acoustic environment and its association with human heart rate: Evidence from real-world data logging with hearing aids and wearables*. <https://doi.org/10.1101/2020.07.08.180869>.
- Crone, E. A., Van Der Veen, F. M., Van Der Molen, M. W., Somsen, R. J. M., Van Beek, B., Jennings, J. R. (2003). Cardiac concomitants of feedback processing. *Biol Psychol, 64*, 143–156.
- Del-Rio-Guerrero, V., Martínez-Martínez, L. A., Arias-Callejas, K., Carbonell-Bobadilla, N., Mejía-Segura, A., Azamar-Morales, G., Espinosa-Orantes, A., Molina-Sánchez, J. R., Mora-Ramírez, M., Mejía-Ávila, M. F., Vargas Guerrero, A., Silveira, L. H., Martínez-Lavín, M. (2023). The value of inquiring patients about local discomfort during blood pressure measurement for fibromyalgia detection. A cross-sectional study. *Semin Arthritis Rheum, 61*, 152218.
- Dickerson, S. S., & Kemeny, M. E. (2004). *Acute stressors and cortisol responses: A theoretical integration and synthesis of laboratory research*. <https://doi.org/10.1037/0033-2909.130.3.355>.
- Diemer, J., Alpers, G. W., Peperkorn, H. M., Shibani, Y., Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: A review of research in virtual reality. *Front Psychol, 6*, 26.
- Ditzen, B., Neumann, I. D., Bodenmann, G., von Dawans, B., Turner, R. A., Ehlert, U., Heinrichs, M. (2007). Effects of different kinds of couple interaction on cortisol and heart rate responses to stress in women. *Psychoneuroendocrinology, 32*, 565–574.
- Draghici, A. E., & Taylor, J. A. (2016). The physiological basis and measurement of heart rate variability in humans. *J Physiol Anthropol, 35*, 22.
- Fairclough, S. H., & Roberts, J. S. (2011). Effects of performance feedback on cardiovascular reactivity and frontal EEG asymmetry. *Int J Psychophysiol, 81*, 291–298.
- Fallon, M. A., Riem, M. M. E., Kunst, L. E., Kop, W. J., Kupper, N. (2021). Multi-modal responses to the virtual reality trier social stress test: A comparison with standard interpersonal and control conditions. *Int J Psychophysiol, 161*, 27–34.
- Forouzanfar, M., Baker, F. C., de Zambotti, M., McCall, C., Giovangrandi, L., Kovacs, G. T. A. (2018). Toward a better noninvasive assessment of preejection period: A novel automatic algorithm for B-point detection and correction on thoracic impedance cardiogram. *Psychophysiology, 55*, e13072.
- Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdiscip Rev Cognit Sci, 11*, e1514.
- Francis, A. L., MacPherson, M. K., Chandrasekaran, B., Alvar, A. M. (2016). Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Front Psychol, 7*, 263.
- Gendolla, G. H. E., & Richter, M. (2006). Cardiovascular reactivity during performance under social observation: The moderating role of task difficulty. *Int J Psychophysiol, 62*, 185–192.
- Gordan, R., Gwathmey, J. K., Xie, L.-H. (2015). Autonomic and endocrine control of cardiovascular function. *World J Cardiol, 7*, 204–214.
- Hruby, F., Álvarez Sánchez, L. F., Ressler, R., Escobar-Briones, E. G. (2020). An empirical study on spatial presence in immersive geo-environments. *J Photogramm Remote Sens Geoinf Sci, 88*, 155–163.
- 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.
- Jaki, T., & Wolfsegger, M. J. (2009). A theoretical framework for estimation of AUCs in complete and incomplete sampling designs. *Stat Biopharm Res, 1*, 176–184.
- 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.
- Jennings, J. R., & Van Der Molen, M. W. (2002). Cardiac timing and the central regulation of action. *Psychol Res, 66*, 337–349.
- Jonas, E., & Mühlberger, C. (2017). Editorial: Social cognition, motivation, and interaction: How do people respond to threats in social interactions? *Front Psychol, 8*, 1577.
- Jönsson, P., Wallergård, M., Österberg, K., Hansen, M., Johansson, G., Karlson, B. (2010). Cardiovascular and cortisol reactivity and habituation to a virtual reality version of the Trier Social Stress Test: A pilot study. *Psychoneuroendocrinology, 35*, 1397–1403.
- Kaur, J., Shivashankarappa, P. G., Sanguida, A., Suganya, M., Ezhumalai, G. (2021). Effectiveness of visual distraction with and without virtual reality glasses in reducing dental anxiety among children with hearing and speech disability: A pilot study. *Int J Clin Pediatr Dent, 14*, S162–S166.
- Keidser, G., Naylor, G., Brungart, D. S., Caduff, A., Campos, J., Carlile, S., Carpenter, M. G., Grimm, G., Hohmann, V., Holube, I., Launer, S., Lunner, T., Mehra, R., Rapport, F., Slaney, M., Smeds, K. (2020). The quest for ecological validity in hearing science: What it is, why it matters, and how to advance it. *Ear Hear, 41*, 5S–19S.
- Koelwijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hear Res, 312*, 114–120.
- Koelwijn, T., Versfeld, N. J., Kramer, S. E. (2017). Effects of attention on the speech reception threshold and pupil response of people with impaired and normal hearing. *Hear Res, 354*, 56–63.
- Koelwijn, 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.
- Kothgassner, O. D., & Felhofer, A. (2020). Does virtual reality help to cut the Gordian knot between ecological validity and experimental control? *Ann Int Commun Assoc, 44*, 210–218.
- Kvavilashvili, L., & Ellis, J. (2004). Ecological validity and the real-life/laboratory controversy in memory research: A critical and historical review. *Hist Philos Psychol, 6*, 59–80.
- Liu, Q., & Zhang, W. (2020). *Sex differences in stress reactivity to the trier social stress test in virtual reality*. <https://doi.org/10.2147/PRBM.S268039>.
- Lovallo, W. R. (2005). *Stress and Health: Biological and Psychological Interactions*. Sage Publications, Inc. <https://doi.org/10.4135/9781452233543>.
- Lozano, D. L., Norman, G., Knox, D., Wood, B. L., Miller, B. D., Emery, C. F., Berntson, G. G. (2007). Where to B in dZ/dt. *Psychophysiology, 44*, 113–119.
- Mackersie, C. L., & Calderon-Moultrie, N. (2016). Autonomic nervous system reactivity during speech repetition tasks: Heart rate variability and skin conductance. *Ear Hear, 37*, 118S–125S.
- Mackersie, C. L., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *J Am Acad Audiol, 22*, 113–122.
- Mackersie, C. L., & Kearney, L. (2017). Autonomic nervous system responses to hearing-related demand and evaluative threat. *Am J Audiol, 26*(3S), 373–377.

- 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.
- Malik, M., Camm, A. J., Bigger, J. T., Breithardt, G., Cerutti, S., Cohen, R. J., Coumel, P., Fallen, E. L., Kennedy, H. L., Kleiger, R. E., Lombardi, F., Malliani, A., Moss, A. J., Rottman, J. N., Schmidt, G., Schwartz, P. J., Singer, D. H. (1996). Heart rate variability. Standards of measurement, physiological interpretation, and clinical use. *Eur Heart J, 17*, 354–381.
- Martínez-Cagigal, V. (2021). *Multiple Testing Toolbox*. MATLAB Central File Exchange.
- Matthen, M. (2016). Effort and displeasure in people who are hard of hearing. *Ear Hear, 37*, 285–345.
- McCorry, L. K. (2007). Physiology of the autonomic nervous system. *Am J Pharm Educ, 71*, 78.
- 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.
- Newlin, D. B., & Levenson, R. W. (1979). Pre-ejection period: Measuring beta-adrenergic influences upon the heart. *Psychophysiology, 16*, 546–553.
- Pichora-Fuller, K. (2016). How social psychological factors may modulate auditory and cognitive functioning during listening. *Ear Hear, 37*, 92S–100S.
- Pichora-Fuller, 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.
- Picou, E. M., & Ricketts, T. A. (2014). Increasing motivation changes subjective reports of listening effort and choice of coping strategy. *Int J Audiol, 53*, 418–426.
- Pielage, H., Plain, B. J., Saunders, G. H., Versfeld, N. J., Lunner, T., Kramer, S. E., Zekveld, A. A. (2023). Copresence was found to be related to some pupil measures in persons with hearing loss while they performed a speech-in-noise task. *Ear Hear*. Retrieved from [https://journals.lww.com/ear-hearing/Fulltext/9900/Copresence\\_Was\\_Found\\_to\\_Be\\_Related\\_to\\_Some\\_Pupil.135.aspx](https://journals.lww.com/ear-hearing/Fulltext/9900/Copresence_Was_Found_to_Be_Related_to_Some_Pupil.135.aspx).
- Pielage, H., Plain, B. J., van de Ven, S., Saunders, G. H., Versfeld, N. J., Kramer, S. E., Zekveld, A. A. (n.d.). Using pupillometry in virtual reality as a tool for speech-in-noise research. *Ear Hear*.
- Pielage, H., Zekveld, A. A., Saunders, G. H., Versfeld, N. J., Lunner, T., Kramer, S. E. (2021). The presence of another individual influences listening effort, but not performance. *Ear Hear, 42*, 1577–1589.
- Plain, B., Pielage, H., Richter, M., Bhuiyan, T. A., Lunner, T., Kramer, S. E., Zekveld, A. A. (2021). Social observation increases the cardiovascular response of hearing-impaired listeners during a speech reception task. *Hear Res, 410*, 108334.
- Plain, B., Richter, M., Zekveld, A. A., Lunner, T., Bhuiyan, T., Kramer, S. E. (2020). Investigating the influences of task demand and reward on cardiac pre-ejection period reactivity during a speech-in-noise task. *Ear Hear, 42*, 718–731.
- Raza, S. B., Patterson, R. P., Wang, L. (1992). Filtering respiration and low-frequency movement artefacts from the cardiogenic electrical impedance signal. *Med Biol Eng Comput, 30*, 556–561.
- Richter, M., Bhuiyan, T., Bramslov, L., Innes-Brown, H., Fiedler, L., Hadley, L. V., Naylor, G., Saunders, G. H., Wendt, D., Whitmer, W. M., Zekveld, A. A., Kramer, S. E. (2023). Combining multiple psychophysiological measures of listening effort: Challenges and recommendations. *Semin Hear, 44*, 95–105.
- Richter, M., Friedrich, A., Gendolla, G. H. E. (2008). Task difficulty effects on cardiac activity. *Psychophysiology, 45*, 869–875.
- Richter, M., & Gendolla, G. H. E. (2009). The heart contracts to reward: Monetary incentives and pre-ejection period. *Psychophysiology, 46*, 451–457.
- Schubert, T., Friedmann, F., Regenbrecht, H. (2001). The experience of presence: Factor analytic insights. *Presence, 10*, 266–281.
- Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *J Speech Lang Hear Res, 58*, 1781–1792.
- Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Front Public Health, 5*, 258.
- Sherwood, A., Allen, M. T., Fahrenberg, J., Kelsey, R. M., Lovallo, W. R., van Doornen, L. J. P. (1990). Methodological guidelines for impedance cardiography. *Psychophysiology, 27*, 1–23.
- Shiban, Y., Diemer, J., Brandl, S., Zack, R., Mühlberger, A., Wüst, S. (2016). Trier Social Stress Test in vivo and in virtual reality: Dissociation of response domains. *Int J Psychophysiol, 110*, 47–55.
- Shields, C., Willis, H., Nichani, J., Sladen, M., Kluk-de Kort, K. (2022). Listening effort: WHAT is it, HOW is it measured and WHY is it important? *Cochlear Implants Int, 23*, 114–117.
- Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends Amplif, 12*, 283–299.
- Shoushtarian, M., Weder, S., Innes-Brown, H., McKay, C. M. (2019). Assessing hearing by measuring heartbeat: The effect of sound level. *PLoS One, 14*, e0212940.
- Slade, K., Kramer, S. E., Fairclough, S., Richter, M. (2021). Effortful listening: Sympathetic activity varies as a function of listening demand but parasympathetic activity does not. *Hear Res, 410*, 108348.
- Smeds, K., Wolters, F., Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *J Am Acad Audiol, 26*, 183–196.
- Somsen, R. J. M., Van Der Molen, M. W., Jennings, J. R., Van Beek, B. (2000). Wisconsin Card Sorting in adolescents: Analysis of performance, response times and heart rate. *Acta Psychol (Amst)*, 104, 227–257.
- Standard, W., Zielasko, D., Von Dawans, B., Domes, G., & Weyers, B. (2020). *Open TSSST VR: Trier social stress test in virtual reality*. Gesellschaft Für Informatik e.V. GI VR/AR Workshop, September.
- Szilágyi, N., Láng, E., Balázs, L. (1992). Computer determination of systolic time intervals based on impedance cardiography. *Int J Psychophysiol, 13*, 45–49.
- Tarvainen, M. P., Niskanen, J. P., Lipponen, J. A., Ranta-aho, P. O., Karjalainen, P. A. (2014). Kubios HRV—Heart rate variability analysis software. *Comput Methods Programs Biomed, 113*, 210–220.
- Van Der Veen, F. M., Van Der Molen, M. W., Crone, E. A., Jennings, J. R. (2004). Phasic heart rate responses to performance feedback in a time production task: Effects of information versus valence. *Biol Psychol, 65*, 147–161.
- van Lien, R., Schutte, N. M., Meijer, J. H., de Geus, E. J. C. (2013). Estimated pre-ejection period (PEP) based on the detection of the R-wave and dZ/dt-min peaks does not adequately reflect the actual PEP across a wide range of laboratory and ambulatory conditions. *Int J Psychophysiol, 87*, 60–69.
- Van Twillert, B., Bremer, M., Faber, A. W. (2007). Computer-generated virtual reality to control pain and anxiety in pediatric and adult burn patients during wound dressing changes. *J Burn Care Res, 28*, 694–702.
- Vasconcelos-Raposo, J., Bessa, M., Melo, M., Barbosa, L., Rodrigues, R., Teixeira, C. M., Cabral, L., Sousa, A. A. (2016). Adaptation and validation of the Igroup presence questionnaire (IPQ) in a Portuguese sample. *Presence, 25*, 191–203.
- 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.
- Vrana, S. R., & Rollock, D. (1998). Physiological response to a minimal social encounter: Effects of gender, ethnicity, and social context. *Psychophysiology, 35*, 462–469.
- Wright, R. A., Dill, J. C., Geen, R. G., Anderson, C. A. (1998). Social evaluation influence on cardiovascular response to a fixed behavioral challenge: Effects across a range of difficulty levels. *Ann Behav Med, 20*, 277–285.
- Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., Oleson, J. (2018). Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. *Ear Hear, 39*, 293–304.
- 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., 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.
- Zimmer, P., Buttlar, B., Halbeisen, G., Walther, E., Domes, G. (2018). Virtually stressed? A refined virtual reality adaptation of the Trier Social Stress Test (TSST) induces robust endocrine responses. *Psychoneuroendocrinology, 101*, 186–192.
- Zimmer, P., Buttlar, B., Halbeisen, G., Walther, E., Domes, G. (2019). Virtually stressed? A refined virtual reality adaptation of the Trier Social Stress Test (TSST) induces robust endocrine responses. *Psychoneuroendocrinology, 101*, 186–192.