A comparison of simultaneously-obtained measures of listening effort: pupil dilation, verbal response time and self-rating

Chiara Visentin[a ,*] , Chiara Valzolgher[b ,c ,*] , Matteo Pellegatti[a] , Paola Potente[b] , Francesco Pavani[b ,c ,d] , and Nicola Prodi[a]

[a] Department of Engineering, University of Ferrara, Ferrara, Italy
[b] Center for Mind/Brain Sciences (CIMeC), University of Trento, Trento, Italy
[c] Centre de Recherche en Neuroscience de Lyon (CRNL), Integrative, Multisensory, Perception, Action and Cognition Team (IMPACT), Lyon, France
[d] Department of Psychology and Cognitive Sciences (DiPSCo), University of Trento, Trento, Italy

* These authors equally contributed to the present work.

Corresponding authors

Chiara Visentin vsnchr@unife.it Department of Engineering, University of Ferrara, Via Saragat 1, Ferrara, 44122, Italy;
Chiara Valzolgher chiara.valzolgher@unitn.it Center for Mind/Brain Sciences (CIMeC), University of Trento, corso Bettini 31, Rovereto (TN), 38068, Italy.

Abstract

OBJECTIVE The aim of this study was to assess to what extent *simultaneously-obtained* measures of listening effort (task-evoked pupil dilation, verbal response time [RT], and self-rating) could be sensitive to auditory and cognitive manipulations in a speech perception task. The study also aimed to explore the possible relationship between RT and pupil dilation.

DESIGN A within-group design was adopted. All participants were administered the Matrix Sentence Test in 12 conditions (signal-to-noise ratios [SNR] of −3, −6, −9 dB; attentional resources focussed *vs* divided; spatial priors present *vs* absent).

STUDY SAMPLE Twenty-four normal-hearing adults, 20–41 years old (*M* = 23.5), were recruited in the study.

RESULTS A significant effect of the SNR was found for all measures. However, pupil dilation discriminated only partially between the SNRs. Neither of the cognitive manipulations were effective in modulating the measures. No relationship emerged between pupil dilation, RT and self-ratings.

CONCLUSIONS RT, pupil dilation, and self-ratings can be obtained simultaneously when administering speech perception tasks, even though some limitations remain related to the absence of a retention period after the listening phase. The sensitivity of the three measures to changes in the auditory environment differs. RTs and self-ratings proved most sensitive to changes in SNR.

Introduction

Listening in noise is a constant challenge. Adverse listening conditions increase the risk of making more speech identification errors (i.e. lower speech intelligibility, SI) than 'ideal' quiet conditions (Mattys et al. 2012). Even when performance is unaffected, the correct identification of speech from a degraded stimulus requires mobilisation of cognitive resources, leading to increased *listening effort*. Within the Framework for Understanding Effortful Listening (FUEL), listening effort is defined as 'the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task, with listening effort applying more specifically when tasks involve listening' (Pichora-Fuller et al. 2016). The definition emphasises that listening effort depends on an additional dimension besides the cognitive demands of a task, and that is the listener's motivation to perform the task. The stronger their motivation, the more listeners will be willing to put effort into the task, regardless of its demands (Peelle 2018).

In the last decade, the construct of listening effort has been recognised as an important dimension in everyday life (Lemke and Besser 2016), and it has become the object of intensive research. The concept is especially relevant for people with hearing loss, who often describe their listening experience as cognitively demanding and tiring (Anderson Gosselin and Gagné 2010). A reliable measure of listening effort for use in clinical practice is still lacking (Pichora-Fuller et al. 2016; McGarrigle et al. 2014). This poses a problem because such a measure could shed light on aspects of hearing impairment impossible to assess with traditional hearing tests (based on pure tone audiometry or accuracy in speech-in-noise tests). It would also be helpful when treating people with hearing loss and/or screening for borderline hearing disabilities (McGarrigle et al. 2014).

It is becoming increasingly evident that different subjective, behavioural and physiological measures of listening effort assess different constructs, or tap into different underlying dimensions of the phenomenon (Strand et al. 2018; Lemke and Besser 2016; McGarrigle et al. 2014). Though often grouped under the umbrella term of 'listening effort measures', they do not necessarily provide converging or overlapping information (Strand et al. 2018). They might possibly give us complementary information on different aspects of this complex construct. In a listening task that involved recalling digits, for instance, Alhanbali and colleagues (2019) simultaneously measured different indices of listening effort (pupil size, electroencephalographic alpha power, skin conductance, and self-ratings). Their results showed weak or non-significant correlations between the measures, and the Authors concluded that different measures should not be used interchangeably. When aiming to assess listening effort in clinical practice, the implication of this finding is that it is crucial to include different types of assessment in the experimental paradigm (Hornsby 2013), and it is more appropriate to combine multiple measures than to use only one (Pichora-Fuller et al. 2016).

In the present study, three proxy measures of listening effort were obtained simultaneously, during a speech perception task, assuming that each measure could convey different information on the underlying construct. We used self-rating as a subjective measure, verbal response time (RT) as a behavioural measure, and task-evoked pupil dilation as a physiological measure. The latter two measures were chosen specifically over other behavioural and physiological measures of listening effort for their potential smooth inclusion in standard speech-in-noise assessments. We adopted a listening test specifically designed for use in the clinical setting, and now translated into multiple languages—the Matrix Sentence Test (Kollmeier et al. 2015). The test was designed to assess speech recognition in background noise by using semantically unpredictable sentences instead of isolated words, non-words or digits. The use of sentences as speech material is much closer to the everyday listening experience, and can thus allow for results that can be more representative of the cognitive process involved in real communication.

Verbal response time and listening effort

Listening effort has been assessed in children and adults using verbal RT in a single-task paradigm (see McGarrigle et al. (2014) and Pichora-Fuller et al. (2016) for reviews), with participants performing a single listening task, during which both accuracy and RTs are recorded. Participants respond to a stimulus by verbally repeating it (verbal RT) or pressing a button (manual RT), and their RT is defined as the time elapsing between the offset of the stimulus and the onset of the participant's response. RT is intended as a measure of processing speed, which is associated with the amount of cognitive capacity allocated to processing the auditory signal (Pichora-Fuller et al. 2016). Slower RTs are thought to reflect an increase in listening effort (Gatehouse and Gordon 1990; Houben et al. 2013).

Compared with the RTs in the more commonly used dual-task paradigms (Gagné et al. 2017), single-task, verbal RTs have proved more sensitive to the effect of SI, in both children (McGarrigle et al. 2019) and adults (Pals et al. 2015). The single-task paradigm overcomes two limitations of dual-task paradigms, relating to the need to allocate attentional resources to only one of the two tasks involved (an aspect especially relevant in children; see Choi et al. 2008), and to the complexity of the secondary task, which can significantly affect the outcomes of the paradigm as a whole (Picou and Ricketts 2014).

In children and adults alike, RT is longer when the signal-to-noise ratio (SNR) decreases (Houben et al. 2013; Mealings et al. 2015; McGarrigle et al. 2019), and the stimulus is more complex (Lewis et al. 2016; Uslar et al. 2013). RTs are sensitive to the type of background noise (Prodi et al. 2019a; Visentin et al. 2019; Meister et al. 2018), to age (Meister et al. 2018, Prodi et al. 2019a, 2019b), and to room acoustics (Prodi and Visentin 2019; Visentin et al. 2018). In all age groups, RTs are shorter when noise reduction methods are adopted (Gustafson et al. 2014; van den Tillaart-Haverkate et al. 2017), for children with normal hearing than for those with hearing loss (McGarrigle et al. 2019), and for children with normal hearing than for deaf children with cochlear implants (Steel et al. 2015). As concerns the effects of auditory manipulations of the SNR in sentence recognition tasks, RTs have been found consistently longer with lower SNRs, whether SI remains constant (Houben et al. 2013; van den Tillaart-Haverkate et al. 2017; Pals et al. 2015) or decreases (Prodi and Visentin 2019; McGarrigle et al. 2019).

In short, using RT as a behavioural measure seems a viable clinical and experimental method for measuring listening effort. It is easy to obtain and can be combined with already routinely performed speech-in-noise tests (Pals et al. 2015; McGarrigle et al. 2019). In a study by Meister et al. (2018), RT was measured using conventional speech audiometry, based on a matrix sentence test, in normal-hearing young listeners, older listeners with clinically normal hearing, and older listeners using hearing aids. RT proved sensitive to the manipulation of SI levels, type of noise (stationary, fluctuating), and listener group. The Authors suggested that RT has the potential to be included in the conventional testing of speech in noise, providing additional information beyond accuracy and self-ratings (Meister et al. 2018). Houben et al. (2013), and Pals et al. (2015) came to similar conclusions using different test materials (that involved identifying the last digit in a triplet, and conversational sentences, respectively).

Pupillometry and listening effort

A commonly used physiological measure of listening effort is task-related pupil dilation (Pichora-Fuller et al. 2016). Pupil size is considered an indicator of cognitive processing load (Kahneman 1973). As long as the listener is engaged in the task, larger task-evoked pupil responses are expected when speech processing is cognitively demanding (Peelle 2018). The task-evoked pupil response is defined as the change in pupil diameter that follows the onset of a momentary auditory event (phasic change; Aston-Jones and Cohen 2005; Zekveld et al. 2018). Pupillometry tracks spontaneous reactions that occur without any need for an explicit response from the participant.

In sentence processing tasks, pupillary responses have proved sensitive to SI levels (Zekveld et al. 2010; Wendt et al. 2018; Zekveld et al. 2014), the degree of auditory spectral resolution (Winn et al. 2015), the type of background noise (Koelewijn et al. 2012), syntactic complexity (Piquado et al. 2010), the SNR (Koelewijn et al. 2014; Lau et al. 2019), and attentional manipulations (Koelewijn et al. 2014, 2015). Pupil dilations capture changes in the listening effort relating to motivation as well as to the cognitive demands of the task in hand (Koelewijn et al. 2018; Ohlenforst et al. 2018; Pichora-Fuller et al. 2016).

As for the effect of auditory manipulations of the SNR on sentence recognition tasks, a greater task-evoked pupil dilation was generally found associated with a decreasing SNR. This effect is mediated by the listener’s motivation, however: when a task becomes too difficult (when SI is too low), listeners simply give up on the task and pupil dilation decreases (Ohlenforst et al. 2018; Wendt et al. 2018; Zekveld et al. 2014). The maximum pupil dilation is generally observed at around 40–50% of SI (Wendt et al. 2018).

Koelewijn et al. (2014) explored the effect of cognitive manipulations on pupil dilation (focussed *vs* divided attention: listening to only one or two speakers at the same time) using a sentence recognition task in fluctuating noise and three SNRs (+3, −3, −9 dB). Their results indicated a worse performance, with increased mean and peak pupil size which was interpreted as increased listening effort when attention was divided than when it was focussed. This is consistent with the general conclusions in the cognitive psychology literature, that typically associate better processing resources (as revealed by a higher accuracy and a lower RT) when participants can focus their selective attention on a given target (or location) compared with when they have to divide it among multiple targets (or locations) (e.g. Driver 2001). The effect of the SNR was only apparent for mean pupil dilation. In a subsequent study, Koelewijn et al. (2015) investigated how attentional processes could use available cues (prior knowledge of target speaker location, target speech onset, or target speaker identity) to facilitate target-masker segregation processes. Their results showed that uncertainty regarding location had a negative effect on performance and resulted in a larger peak pupil dilation (interpreted by the Authors as an increase in cognitive load), thus confirming previous findings regarding the essential role of selective attention (Koelewijn et al. 2014). It has also been demonstrated that knowing ‘where’ to listen makes it easier to segregate the target in adverse listening situations (Best 2007).

One crucial advantage of pupillometry over behavioural or subjective measures of listening effort is that pupil size varies *during* the task, continuously tracking changes in cognitive resource allocation over time (Winn et al. 2018). Conversely, behavioural and self-report measures reflect changes that occur after the speech processing phase (Peelle 2018). To track changes in pupil dilation effectively over time, it is therefore important to consider changes in pupil morphology as a function of time rather than as a mean value. This can be done using time series, growth curve analyses (Wendt et al. 2018; McGarrigle et al. 2017a; McGarrigle et al. 2017b; Wagner et al. 2019), or analyses on time windows tailored to a given study design (Winn et al. 2015; Winn 2016; Wendt et al. 2016). To give an example, changes in task-evoked pupil dilation due to external factors were tracked effectively by adopting time windows that primarily covered the listening phase, from stimulus onset to peak pupil dilation 500 ms after stimulus offset (Winn et al. 2015; McGarrigle et al. 2017), or even by splitting the listening phase into two different epochs (Wendt et al. 2016).

Simultaneous measures of response time and pupil dilation

Pupil dilation and RT have rarely been measured simultaneously (i.e. obtained within the same trial). The reason for this is primarily methodological: RT should be measured immediately after stimulus offset, to obtain information only on the amount of cognitive capacity allocated to processing the auditory stimulus and avoid contaminating the measure with processing times related to memory components (Alhanbali et al. 2019). On the contrary, a fixed amount of time between stimulus offset and response prompt (retention period) is usually required for pupillometry (Winn et al. 2018). Task-related pupil response is slow, with latencies in the maximum response of the order of several hundred milliseconds. In sentence recognition experiments, Winn et al. (2018) suggested that pupil will start to dilate roughly 0.5–1.3 s after stimulus onset and dilation will peak roughly 700 ms–1 s after stimulus offset, within the retention period. This period is useful to avoid stimulus-evoked pupil dilations being contaminated by pupil dilations related to response preparation and delivery (McCloy et al. 2016). In the case of behavioural responses given by pressing a button, it has been demonstrated that 70% of the pupil response could be attributed to response preparation, starting as early as 400 ms prior to pressing the button (Hupé et al. 2009). Using a shorter stimulus-to-prompt delay (1.5 *vs* 3 s), Winn (2016) found pupil dilations three times greater than those seen in a study by Zekveld et al. (2010) for a comparable speech quality. The Authors concluded that the stimulus response timing prompt influences the extent of pupillary response.

To the best of our knowledge, RTs and pupil dilation data were only collected simultaneously in three studies. Steel et al. (2015) administered a two-alternative forced choice test designed to assess binaural fusion to deaf children with cochlear implants and children with normal hearing. The percentage change in pupillary diameter was obtained during the task, and the peak pupil diameter was calculated during the 2 s following the stimulus onset. Manual RTs were recorded as the time elapsing from stimulus onset to manual response and correlated with the percent change in pupillary diameter relative to baseline values. Both measures were higher when binaural fusion was lower, in both groups of participants. A significant positive correlation (R = 0.69) was found between the two measures, and the Authors suggested that they might both reflect much the same cognitive processing. In the other study, McGarrigle et al. (2017a) examined the effect of SNR (+15, −2 dB) on performance in a speech-picture verification task administered to 8- to 11-year-olds. Task-evoked pupil dilation was analysed over a period of 2.5 s, starting at stimulus onset. This time was chosen because it included pupil dilation from stimulus onset to peak response, but included only part of the stimulus presentation (the shortest stimulus lasted 13 s). The results showed no differences in RT between listening conditions, whereas participants had a larger mean pupil dilation in the unfavourable than in the favourable listening condition, which was interpreted as a physiological indicator of increased listening effort. No significant relationship emerged between the two measures. Finally, McGarrigle et al. (2017b) examined changes in behavioural (RT) and physiological (pupil size) indices of listening-related fatigue during the same listening task as presented in McGarrigle et al. (2017a), administered to adults in two SNRs (+15, −8 dB). The results indicated that (manual) RT did not reveal any difference between the listening conditions, while changes in pupil size occurred both in response to the SNR and trial position.

This study

The first aim of this study was to confirm the significant effect of an auditory manipulation (SNR) in a speech perception task on physiological, behavioural and subjective measures of listening effort, in the specific case the measures were *simultaneously-obtained*. Task-evoked pupil dilation was used as a physiological indicator, measured over the period of stimulus presentation. Behavioural measures were obtained using the verbal RT, and subjective measures using self-ratings on a visual-analog scale (VAS). Previous studies where the measures were acquired individually suggested that the SNR should have a significant impact on all three measures of listening effort. Therefore we expected a worse SNR to coincide with a longer RT latency (Prodi and Visentin 2019), larger pupil dilation (Koelewijn et al. 2014), and greater self-rated listening effort (Lau et al. 2019).

A second aim of the present study was to examine how physiological, behavioural, and subjective measures of listening effort relate to one another, and how the pattern of correlations differed depending on the SNR. While the link between different measures of listening effort obtained simultaneously has been explored in previous studies (e.g. Strand et al. 2018; Alhanbali et al. 2019), the specific relation between RT and task-evoked pupil response remained largely unexplored in the literature. The results of the present study will add to the current literature on the topic, by exploring this relation in the adult population. Compared to the study by McGarrigle et al. (2017b), the current study will examine the relationship using shorter speech material (as generally used in literature studies on listening effort) and a more demanding task (word-to-word articulation of the responses). For the present study, we opted not to include a retention period after the sentence playback in the experimental paradigm. The choice was motivated by the necessity of obtaining an accurate measure of RT. Given the evidence on the influence of response preparation and delivery on pupil dilation, we decided to analyse the pupil trace over a time window corresponding to the listening phase alone. The choice was conservative, to avoid the inclusion of pupillary responses unrelated to task-evoked dilation in the analyses. A time series analysis was performed to elucidate variations in the shape of the pupil dilation time course, rather than averaging all the data to obtain a single value.

A third aim of the study was to see how sensitive the three measures are to two types of cognitive manipulation: a change in the direction of the attentional resources (fixed *vs*. random); and a change in the spatial priors regarding the position of the target auditory stream, as conveyed through minimal visual cues (i.e. visual place-holders indicating the actual position of the speakers hidden from view behind a white curtain). While previous works have examined the impact of these cognitive variables while listening in noise (e.g. Koelewijn et al. 2015; Best et al. 2007) it remains to be ascertained to what extent the various physiological, behavioural and subjective measures of listening effort may be *differently affected* by these manipulations. For instance, Koelewijn et al. (2015) showed that participants have a greater task-evoked pupil dilation and greater listening effort in the random attention condition (with a random speaker location) compared with the fixed attention condition (fixed speaker location), but did not examine the effect on RTs. Instead, Best et al. (2007) documented improved performance related to the provision of visual cues (i.e. marking the location of the source), but did not address the effect of this visual information on listening effort or pupillometry. With the here proposed cognitive manipulation, we introduced the idea of having sound sources in the external space. We did not ask participants to wear headphones. Conversely, we moved the sources in the real world as external objects. This strategy gave us the opportunity to study the impact of having visual references about the real spatial positions of the two sources involved compared to a situation in which it was impossible to locate precisely the sources in the external space. We aimed to replicate previous finding but modifying the perspective of the acoustic experience creating a more complex scenario in which the sources are external and spatially identifiable. This choice was dictated by the belief that acoustic space and the experience of listening in noise are part of a single complex process.

Materials and methods

Participants

Twenty-four participants (17 females) took part in the study at the University of Trento (age: *M* = 23.5, *SD* = 5.2, range = 20–41 years). The sample size was first calculated based on the study by Koelewijn et al. (2014) (*d* = 0.72, with a SNR difference of 6 dB). Power analysis revealed that for a significance level of 0.05 and a power of 80% a sample size of 14 was required (two-tailed paired *t* test). Due to the need to balance 24 conditions (three SNR, two attention conditions, and two visual conditions) the sample size was then extended to 24 participants.

All participants reported having no history of auditory or neurological disease and their vision was normal or corrected to normal (with contact lenses). To ensure normal hearing, we measured pure tone hearing thresholds with an audiometer (Grason Stadler GSI 17) at the frequencies 250, 500, 1000, 2000, and 4000 kHz for both right and left ear for each participant prior to the experiment (as in Koelewijn et al. 2014). All subjects had normal hearing, defined as thresholds less than or equal to 20 dB HL at these frequencies for both ears.

Participants all signed an informed consent form before starting the experiment, which was conducted in accordance with the Declaration of Helsinki (1964, amended in 2013) and the ethical regulations at the University of Trento.

Apparatus

Participants sat in a sound-proofed and partially anechoic booth (Amplifon G2x2.5; floor area = 200 × 250 cm, height = 220 cm), 60 cm away from a white fabric curtain, with the position of their head stabilised using a chin rest. Participants were asked to fixate on a green dot produced by a LED (diameter: 0.50 cm; elevation: 40 cm), located approximately at eye level, in the middle of the curtain (Figure 1).

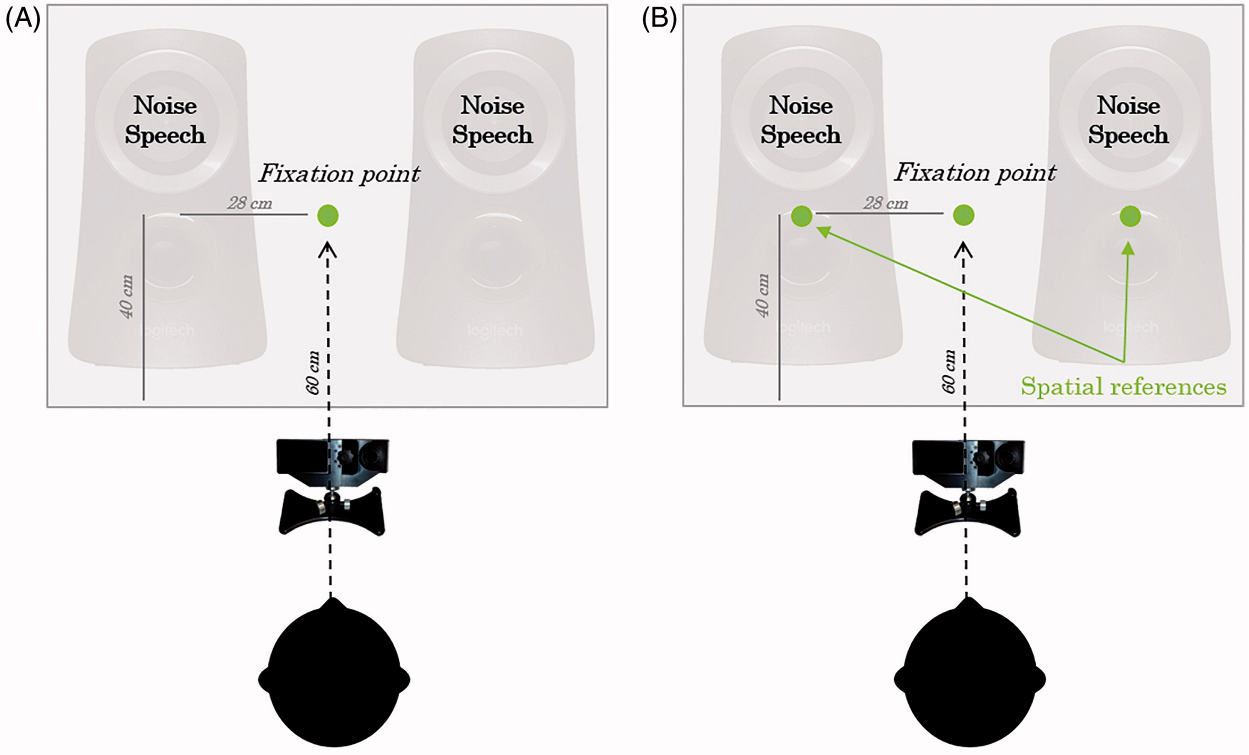

Figure 1. Experimental setup for the sessions without (A) and with (B) a visual reference. The loudspeakers were placed with the centre 28 cm away from the fixation LED (corresponding to 25° and 335° azimuth).

Auditory stimuli were presented through two loudspeakers placed on the table on either side of the fixation LED (with the centre 28 cm away from the LED, corresponding to 25° and 335° azimuth), hidden behind the curtain.[1] Two green LEDs served to mark the centre of each loudspeaker's cone when lit (Figure 1(B)). A light source illuminating the curtain was placed 2 m behind the participant. During the task, pupil diameter in the left eye was measured with an infra-red eye tracker (Eyelink 1000 Desktop, Host PC—EMP 300 W; SN AEP08K0011) placed on a table between the participant and the curtain (60 cm away from the participant) at 0° azimuth. Pupillary data were recorded at a sampling rate of 1000 kHz.

Stimuli presentation and data collection were managed by a desktop PC (HP Compaq LA2405X) running a LabVIEW script (version 18.0, 2018) developed in house. The script used MIDI commands to control an audio rendering engine consisting of the AudioMulch software (version 1.0, 2006) with the X-Volver plug-ins. All signals were delivered through a RME Fireface 400 sound card. A microphone was placed on the table 60 cm away from the participant's mouth and connected to the sound card to record the participant's responses.

Speech perception task

Speech perception was assessed with the Italian Matrix Sentence Test (ITAMatrix; Puglisi et al. 2015). The speech stimuli in the ITAMatrix are five-word sentences with a fixed syntactic structure but no semantic predictability (e.g. *Sofia trascina poche matite utili* [Sophie drags few useful pencils]). All sentences were generated using 50 words in very common use as listeners' familiarity with the words minimises the influence of their linguistic competence on their speech perception. Digital recordings of the ITAMatrix sentences were acquired by agreement with the producer (Hoertech GmbH). Sentence delivery took an average 2.3 s. For our experiment, 276 sentences were randomly selected from among the test corpus and organised into 12 lists of 20 sentences, plus two additional lists of 18 sentences for the practice trials. Therefore, each experimental condition was evaluated over 20 trials (i.e. sentences).

For each trial, participants listened to one five-word sentence and, after the audio offset, they repeated as many words of the sentence as they could recall, in sequential order. A score was awarded for each word correctly recalled, and used to calculate the SI as the percentage of correct words in the sentence. At each trial, the RT—defined as the time elapsing between the audio offset and the onset of the participant's response—was also recorded.

Listening conditions

The masker was a stationary noise with the same long-term spectrum as the spoken sentences (Puglisi et al. 2015). It had a fixed sound pressure level of 63 dB(A) at the listener's position, while that of the sentences was varied to achieve three SNRs (−3, −6 and −9 dB). The choice of these SNRs was prompted by the performance of six participants in a pilot study (three females; mean age = 26 years), in which the speech reception thresholds for correct word identification rates of 20%, 50% and 80% were ascertained using an adaptive procedure. The range of SNRs was then chosen, aiming to measure SI values higher than 50% but still below ceiling. Speech and noise levels were obtained from the energetic average of the signals at the two ears of a B&K Type 4100 head and torso simulator in the listener's position.

Depending on the experimental condition, the loudspeaker transmitting the target sentence was either always the one on the left (fixed attention condition) or it changed randomly on a trial-by-trial basis (random attention condition, with 50% of the trials presented on each side). The background noise was always transmitted by both loudspeakers. The duration of each trial was fixed at 5.5 s. The trial started with background noise alone, and the sentence began nearly 3.2 s later. The background noise and the sentence ended simultaneously.

Design and procedure

The experimental protocol was the same for all participants, and consisted of two sessions, separated by nearly a week. At one session, participants performed the task without any visual references, identifying the source of the target sentence using their hearing alone (see Figure 1(A)). At the other, the positions in space of the two loudspeakers (i.e. the two possible sources of the target sentence) were clearly indicated using two visible LEDs (see Figure 1(B)). The order of the two sessions (with and without this visual reference) was counterbalanced across participants. The pause between the two sessions was intended to attenuate the memory of the spatial position of the loudspeakers in the participants. When visual references were provided (see Figure 1(B)), participants were informed that the green LEDs showing up on the white curtain indicated the exact location of the two loudspeakers placed behind the curtain. The LEDs remained on at all times during a trial, so they provided no cues as to a sentence's onset or the side of its delivery. They did offer a clear visual anchor for the participants' attention during sentence presentation, however. Participants were asked to keep their gaze fixed on the central LED throughout the task (i.e. no overt attention orienting was allowed while they were listening).

Each session included an initial practice followed by six experimental blocks. Each one had a fixed SNR (−3, −6, and −9 dB) and a specified attention condition (fixed or random). The resulting design consisted of 3 × 2 blocks (e.g. −3 random, −3 fixed; −6 random, −6 fixed; −9 random, −9 fixed). Each block included 20 trials and the order of blocks was counterbalanced between participants. Following the completion of each block, participants were asked to report how much effort it took to understand the sentences ('*How much effort did listening and understanding these word sequences require?'*). Their answers were given verbally using a 9-point rating scale (from 1 for minimum effort to 9 for maximum effort). Pupil dilation was measured during each trial. A diagram of the experimental procedure is showed in Figure 2.

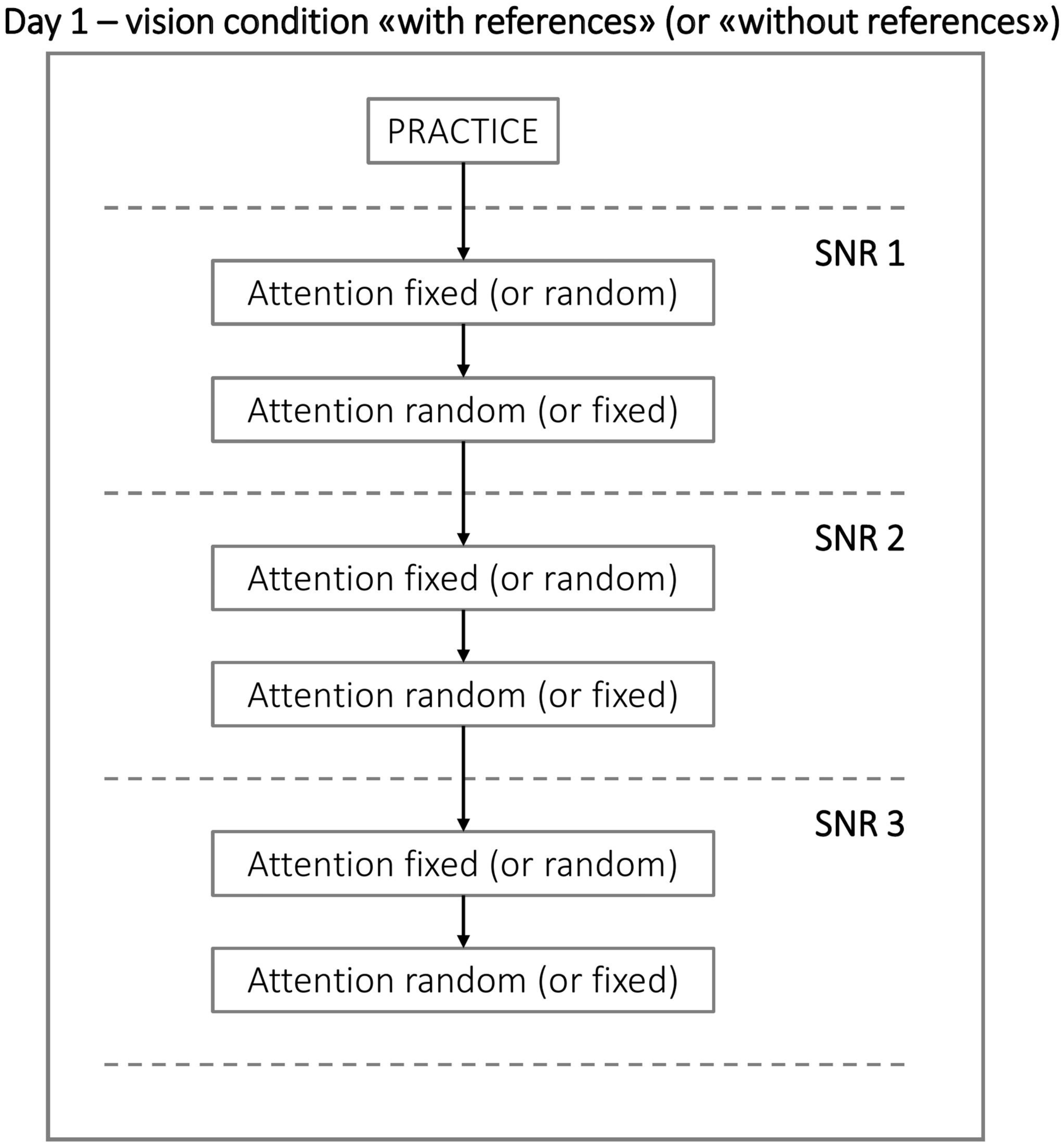
Day 1 – vision condition «with references» (or «without references»)
PRACTICE
SNR 1
Attention fixed (or random)
Attention random (or fixed)
SNR 2
Attention fixed (or random)
Attention random (or fixed)
SNR 3
Attention fixed (or random)
Attention random (or fixed)

Figure 2. Outline of the experimental procedure for one session. Each session was presented in one of the two vision conditions: (i) with visual reference, (ii) without visual references. In each session, the order of the SNRs (−3, −6 and −9 dB) and the order of the attention conditions (random or fixed) was counterbalanced across participants.

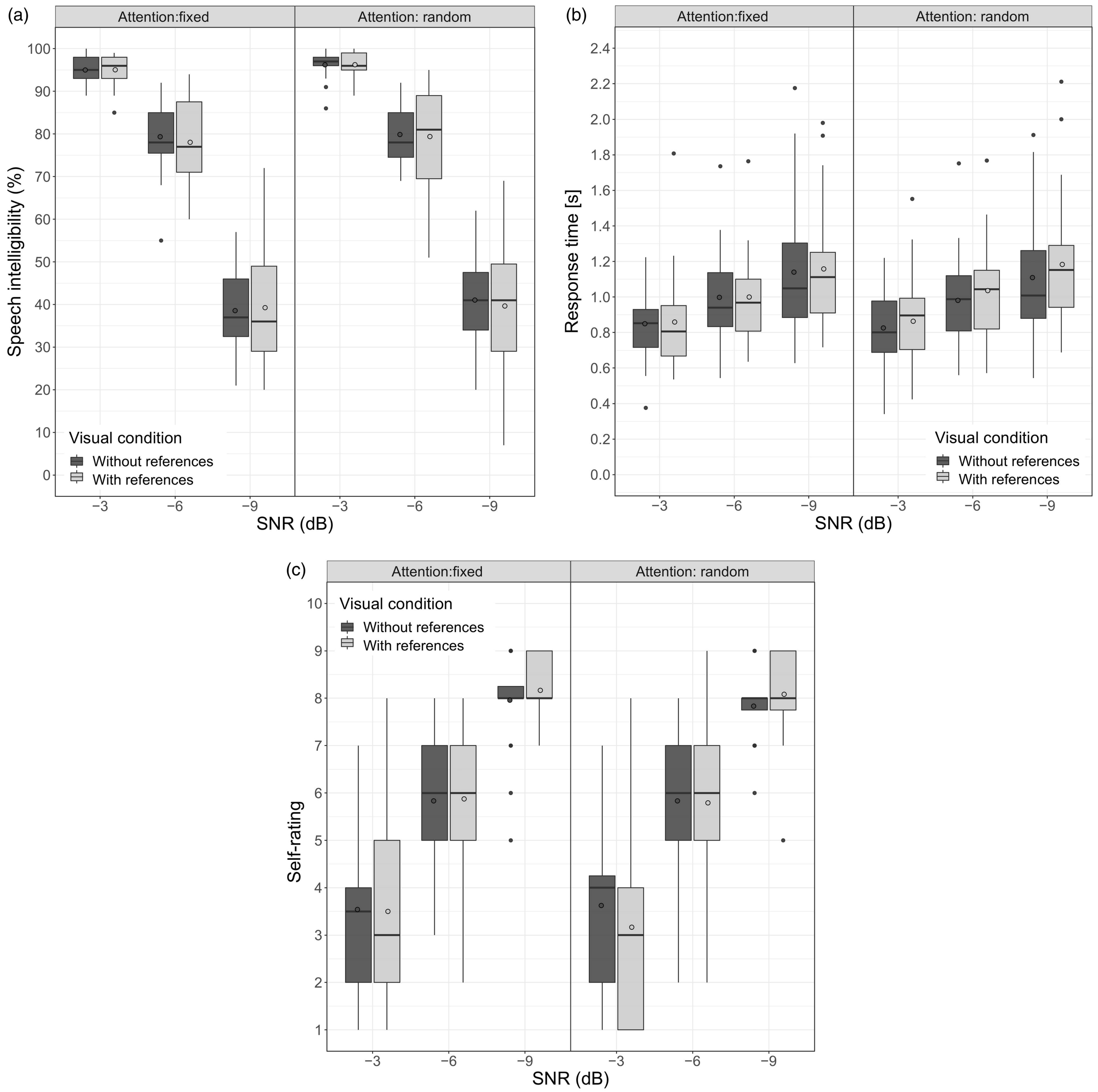

Figure 3. Box plots for: (a) speech intelligibility; (b) response times; and (c) self-ratings of listening effort, by signal-to-noise ratio (SNR; on the *x*-axis), attention conditions (fixed – left panel; random – right panel) and visual conditions (without references – dark grey boxes; with references – light grey boxes). Box plots represent the median (middle line), mean (white circle) and interquartile range of the data distribution; outliers are shown as black circles outside the whiskers.

In the practice at the start of each experimental session, listeners were familiarised with the task using a list of 18 sentences played back in noise. The SNR (0, −6 dB) and the target position (fixed, random) varied between sentences to make participants familiar with the test stimuli (by using a more favourable SNR of 0 dB) and with one of the listening conditions involved in the experiment (SNR = −6 dB). Before starting the practice of the first experimental session, participants' hearing thresholds were measured with an audiometer. The whole procedure, including this hearing threshold measurement, fitting the eye-tracker, the practice and the actual experiment (with breaks) took approximately an hour and 15 min for the first session, and an hour for the second.

Data analysis

Speech intelligibility (SI)

A Generalised Linear Mixed-effects Model (GLMM) was run to identify significant differences in SI across listening conditions. A binomial distribution was adopted for the analysis, given the binary nature of the SI data (0/1 for wrong/right responses at word level, and data bound in the [0;1] interval at sentence level) (Jaeger 2008). The analysis included SNR (−3, −6, −9 dB), attentional condition (random *vs* fixed), and visual condition (with or without a visual reference) as fixed factors; all two- and three-way interactions were considered. The random effect structure accounted for individual variance in intercepts and slopes for the three within-subject factors. Individual intercepts were thus allowed to vary, as well as the effect of each experimental variable on the individual response.[2]

The GLMM was implemented using *R* (R Development Core Team, 2013) and the *lme4* package (Bates et al. 2015). Post-hoc comparisons were run using least-squares means tests (*R* package: *emmeans*; Lenth 2019). The Bonferroni method was applied to adjust the *p*-values in the case of multiple comparisons. The statistical significance threshold was set to $\alpha = 0.05$.

Response time (RT)

Sentence playback and participant's verbal response in each trial were recorded using a microphone located close to the participant's mouth. The audio recordings were analysed using Praat (Boersma and Weenink 2019), based on a visual examination of the time waveform and the automatic detection of silent intervals.

Shapiro-Wilk tests showed that the RT data were not normally distributed ($p < 0.05$). RT data are known to follow a skewed distribution, rising rapidly on the left with a long tail on the right (Whelan 2008). The same GLMM statistical method as for SI was therefore used to examine the effect of the listening conditions on RT too, and a Gamma distribution with a logarithmic link function was adopted to reproduce the characteristics of the raw RT data. The same fixed effects as those described above were included in the GLMM. The initial statistical model included a maximal random effect structure, which did not converge due to over-parametrization. Then random effect terms were systematically removed until the model finally converged. The final random effect structure included participant (random intercept) and SNR (random slope).[3]

Self-ratings of effort

Participants rated their listening effort on a rating scale. A cumulative link mixed model was used for the data analysis (*R* package: *ordinal*; Christensen 2019) because it enables the relation between an ordinal response variable and the independent factors to be described.[4]

Pupil dilation data preprocessing

The recorded data were analysed for 23 of the 24 participants, while for one they were deemed unreliable due to technical issues. Before any analysis, data were processed to convert pupil size from arbitrary camera units to absolute units (mm). An artificial pupil, with a known pupil size (seven different diameters, from 2 to 7 mm), was measured with the eye-tracker. From these calibration measures a mean index of correction was derived. Then eye blinks and saccades (automatically coded by the eye-tracker) were removed from each recorded pupil dilation trace using a linear interpolation, which started 100 samples before and ended 100 samples after the blink/saccade. Following the recommendation in Winn et al. (2018), suggesting that traces with a percentage of blinks (>15–25%) should not be used for further analysis, traces in which more than 25% of the data consisted of blinks or saccades (22.0% of the dataset) were excluded. Blocks for which less than 16 valid traces remained were also removed. After data cleansing, more than half of the experimental blocks for five participants had been removed, so the analyses on the pupil dilation data were run for the remaining group of 18 participants.

All traces were then time-aligned at the beginning of the sentence and baseline-corrected by subtracting the mean pupil size during the 1 s period prior to sentence onset from the value of each time point within a given trace.[5] After baseline correction, traces were averaged for each condition, and the resulting time series were down-sampled to obtain a sample mean pupil size for every 25 ms of the analysis, for each participant, and in each condition. Finally, the pupil dilation traces were cut to retain only the data within the period of time [0; 1.85 s] corresponding to the duration of the shortest sentence. The choice of time window was conservative to ensure that it would only track stimulus related-effort and not be affected by motor planning or the delivery of the verbal response (Winn et al. 2015).

Growth curve analysis

Growth Curve Analysis (GCA) is a statistical method that enables the shape of the curve for pupil response to be analysed by modelling pupil dilation as a function of time (Winn et al. 2018). Analysing the pupil dilation time course is judged to be more effective than traditional approaches as it generates a systematic description of the pupil data instead of only analysing peak or mean pupil dilation over a given period of time (van Rij et al. 2019). This method based on the time course was used in previous studies on listening effort to investigate: the effect of the SNR and masker type on speech processing in adults (Wendt et al. 2018); the effect of the SNR on children (McGarrigle et al. 2017a); and differences between adults with normal hearing and those with hearing impairments in an auditory decision task (Wagner et al. 2019).

GCA is a multi-level regression technique that models changes in pupil dilation over time using orthogonal polynomials, and quantifies the differences between conditions and between participants (Mirman 2014). In the present study, GCA was implemented using *R* (R Development Core Team, 2013), with the *lme4* package (Bates et al. 2015). The visual examination of the pupil trace showed that, in the restricted time window chosen for the analysis, the curve had only one change of direction (i.e. initial change from flatness; Mirman 2014). Therefore, the pupil dilation trace was modelled as a first-order polynomial, thus describing the time course of pupil dilation using two terms: (i) overall average (or 'area under the curve'); and (ii) overall slope of the curve. The SNR (categorical variable: −3, −6, −9 dB), attention condition (fixed, random), and visual modality (with or without a visual reference) were also considered as fixed effects. The model also included a random effect structure capturing variability in the pupil dilation time course at participant level, as well as individual differences in sensitivity to the experimental manipulations.[6] The *p* values for the GLMMs were obtained using likelihood ratio tests. In the case of significant results including the SNR (categorical variable with three levels), multiple pairwise comparisons were conducted with the *multcomp* package (Hothorn et al. 2008). The Bonferroni method was applied to adjust the *p*-values for multiple comparisons.

Correlation analysis

A correlation analysis was run between individual SI scores, RTs, self-ratings and (overall mean) pupil dilation. Correlations were firstly examined using standard Spearman's correlation tests, aggregating data both across all conditions and across the three SNRs. Moreover, following the approach of McGarrigle et al. (2021), a repeated measures correlation was applied to the data. This statistical method examines the overall intra-individual association between two measures (Bakdash and Marusich 2017). It takes into account the non-independence between the data, yielding a greater power than standard regression methods in which data are averaged to meet the assumption of independence. The repeated-measures correlation can detect associations between variables that might otherwise be obscured by artefacts due to aggregation. The main advantages of this regression technique over standard ones are its high statistical power (allowing to test within-subject associations between measures without requiring large samples of participants; McGarrigle et al. 2021) and the possibility to analyse paired repeated measures without averaging or violating independence assumptions (Bakdash and Marusich 2017).

In this study the repeated-measures correlation was used to examine to what extent two measures of effort (e.g. RT and pupil dilation) show corresponding variance as a function of changes in the within-subject factor (SNR). The analysis was implemented using the *rmcorr* package in *R* (Bakdash and Marusich 2017). For both analyses, the Bonferroni method was applied to adjust the *p*-values for multiple comparisons.

Results

Speech intelligibility

The findings for SI are shown in Figure 3(a). Our analysis showed a significant main effect of the SNR ($\chi^2(2) = 1252.71$, $p < 0.001$). The main effects of the attention ($p = 0.06$) and visual condition ($p = 0.89$), the SNR X attention interaction ($p = 0.09$), the SNR X vision interaction ($p = 0.85$), the attention X vision interaction ($p = 0.71$) and the three-way interaction ($p = 0.47$) were not significant. The difference in SI between the conditions without and with visual references was 0.4 percentage points; the difference in SI between the conditions with random and fixed attention was 1.2 percentage points.

Concerning the main effect of the SNR, pairwise comparisons indicated that, when collapsed across attention and visual conditions, SI increased significantly with higher SNRs. In particular, SI rose by 39.0 percentage points for the listening condition with a SNR of −6 as opposed to −9 dB ($-9 < -6$: $z = -22.9$, $p < 0.001$), and by 17.0 percentage points for the listening condition with a SNR of −3 as opposed to −6 dB ($-6 < -3$: $z = -18.51$, $p < 0.001$).

Response time

Our findings for RT are shown in Figure 3(b). The analysis revealed a significant main effect of the SNR ($\chi^2(2) = 70.23$, $p < 0.001$). The main effects of the attention ($p = 0.23$) and visual condition ($p = 0.46$), the SNR X attention interaction ($p = 0.46$), the SNR X vision interaction ($p = 0.052$), the attention X vision interaction ($p = 0.14$) and the three-way interaction ($p = 0.81$) were not significant. The mean difference in RT between the conditions with and without visual references was 33 ms; the difference in RT between the conditions with fixed and random attention was 2 ms.

As concerns the main effect of the SNR, pairwise comparisons showed that, when collapsed across the attention and visual conditions, participants had significantly faster

RTs in more favourable SNRs (−9 > −6: z = 5.74, p < 0.001, ΔRT = 144 ms; −6 > −3: z = 5.67, p < 0.001, ΔRT = 154 ms).

Self-ratings

Participants' ratings of their own listening effort are shown in Figure 3(c). The analysis identified a significant main effect of the SNR (p < 0.001). The main effects of the attention (p = 0.33) and visual condition (p = 0.43), the SNR X attention interaction (p = 0.69), the SNR X vision interaction (p = 0.13), the attention X vision interaction (p = 0.60) and the three-way interaction (p = 0.69) were not significant. The mean difference in self-ratings between the conditions with and without visual references was 0.01; the difference in self-ratings between the conditions with fixed and random attention was 0.09.

The main effect of the SNR showed that, when collapsed across attention and visual conditions, higher self-ratings (indicating that listening was more effortful) were more likely for lower SNRs. Post hoc tests confirmed that self-ratings were higher for the listening condition with the SNR at −9 than when it was −6 dB (z = 7.41, p < 0.001, mean difference: 2.2), and likewise for the SNR at −6 as opposed to −3 dB (z = 5.60, p < 0.001, mean difference: 2.4).

Pupil dilation data

Figure 4 shows the time courses of the mean pupil response averaged across participants by SNR, over the time window [0; 4 s] (with 0 s corresponding to the sentence onset). The pupil trace for the −9 dB SNR condition showed a peak almost 3 s after the sentence onset. Conversely, the traces for the −3 and −6 dB SNR conditions followed a similar, always increasing trend and did not follow the typical pattern of task-evoked pupillometric response reported in literature, due to the contamination with the pupillometric response related to response preparation and delivery.

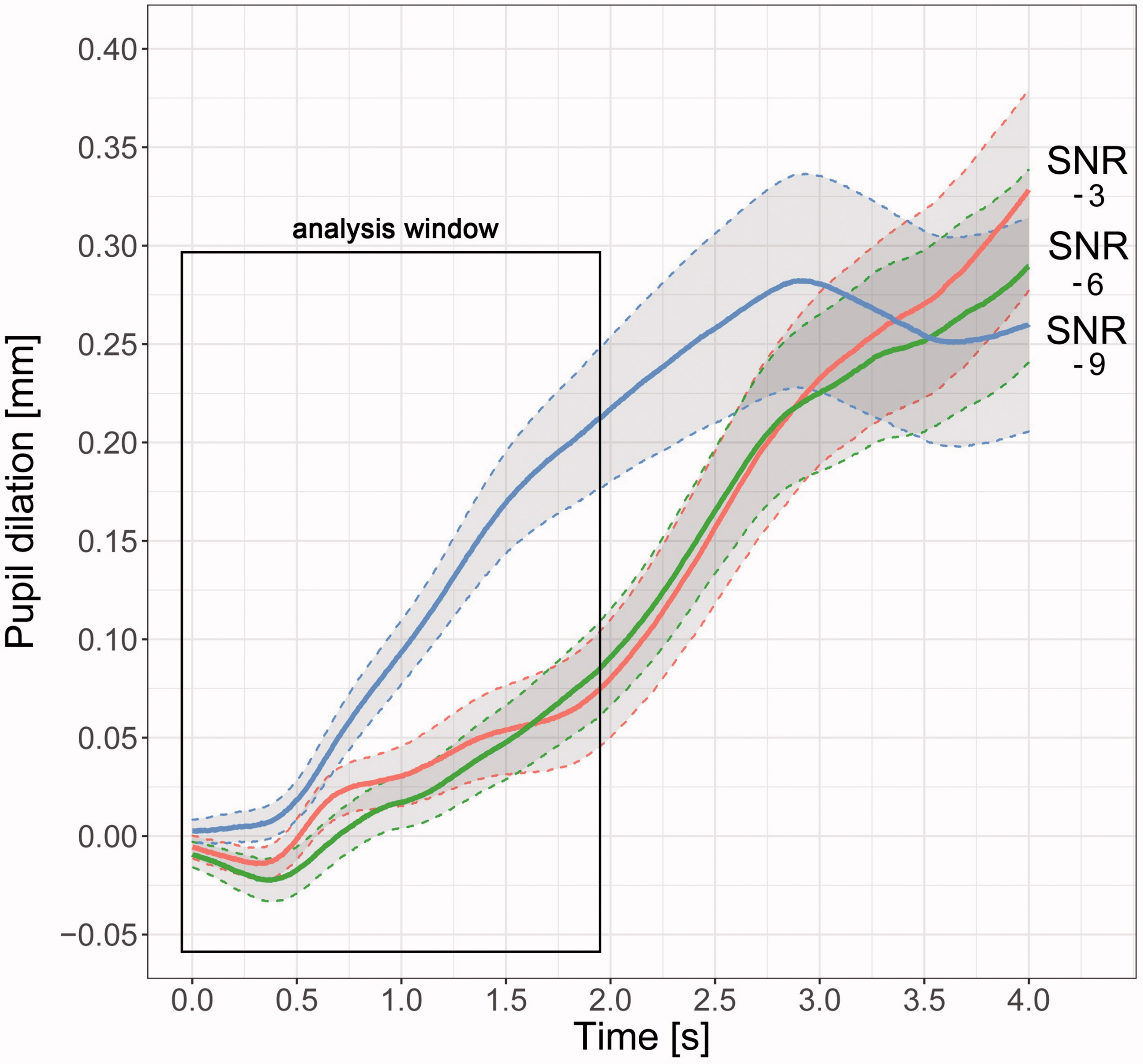

Figure 4. Baseline-adjusted mean pupil dilation over time by SNR (± standard error of the means): –3, –6, –9 dB. The framed period of time refers to the time elapsing between sentence onset (0 s) and the end of the shortest sentence (1.86 s).

With reference to the listening phase ([0; 1.86 s]), GCA revealed a significant effect of the SNR on both polynomial terms (intercept: $\chi^2(2) = 6.03$, $p = 0.042$; linear slope: $\chi^2(2) = 21.01$, $p < 0.001$). There was no difference between the attention ($p = 0.50$ and $p = 0.96$) and visual conditions ($p = 0.27$ and $p = 0.60$) in either of the polynomial terms.[7]

From Figure 4 it can be seen that the course of pupil dilation in the three conditions differed mainly in terms of growth rate. Pairwise comparisons indicated that participants' mean pupil dilation was greater in the −9 dB than in the −6 or −3 dB SNR conditions (−9 > −6: $\beta = 0.054$, $z = 2.98$, $p = 0.018$; −9 > −3: $\beta = 0.047$, $z = 2.59$, $p = 0.028$). Likewise, the linear slope of the pupil dilation trace was steeper in the −9 than in the −6 or −3 dB

SNR condition (−9 > −6: $\beta = 0.22$, $z = 2.39$, $p = 0.048$; −9 > −3: $\beta = 0.23$, $z = 2.46$, $p = 0.028$). No such differences in overall pupil dilation or linear slope emerged between the −3 and −6 dB SNR conditions ($p = 0.74$ and $p = 0.76$, respectively).

Correlations

Standard Spearman's correlation tests were run to examine the relationship between SI scores, RTs, self-rating and pupil dilation. The analyses were conducted on data collapsed across all conditions. Moreover, the association between the measures was explored as a function of the SNR condition, being the only manipulation showing a significant effect on all the measures included in the study. No significant correlation was found between overall measures (all *p*s > 0.43). When collapsing data across attention and visual conditions, significant correlations were found between SI scores and RTs ($\rho = -0.44$, $p = 0.007$), between SI scores and self-ratings ($\rho = -0.77$, $p < 0.001$), and between RTs and self-ratings ($\rho = 0.44$, $p = 0.007$). Mean pupil dilation showed no correlation with SI scores (ρ = −0.37, $p = 0.087$), RTs ($\rho = 0.15$, $p = 0.99$), or self-ratings ($\rho = 0.35$, $p = 0.062$).

A repeated measures correlation analysis was also applied to the data, in order to explore the within-subject association between the measures. The relationships were examined as a function of the SNR condition both overall (i.e. three block values for each participant, with data collapsed across attention and visual conditions), and within each combination of attention and visual conditions. Table 1 shows the results of the rmcorr analysis. Both overall and condition specific analysis showed a negative relationship between SI scores and RTs (with higher scores associated with shorter RTs) and between SI scores and self-ratings (with higher SI related to lower ratings of listening effort). Moreover, changes in RTs as a function of SNRs always showed a positive correlation with changes in self-ratings, with longer RTs associated with higher ratings of listening effort. Pupil dilation showed a significant negative relationship with SI only when overall results were considered, not in condition analyses. Smaller pupil dilations associated with higher SI scores. The relationships between RTs and pupil dilation, and between self-ratings and pupil dilation were always nonsignificant.

Table 1. Results of repeated-measures correlation analysis (*r* and 95% confidence intervals) between the outcome measures. (Table view)

	RTs	Self-ratings	Pupil dilation
Overall			
Speech intelligibility (SI) scores	**−0.77** [−0.88, −0.59]	**−0.89** [−0.94, −0.79]	**−0.52** [−0.73, −0.23]
Response times (RT)		**0.76** [0.58, 0.87]	0.40 [0.08, 0.65]
Self-ratings			0.43 [0.11, 0.67]
Vision: without references, Attention: fixed			
Speech intelligibility (SI) score	**−0.62** [−0.77, −0.40]	**−0.87** [−0.93, −0.78]	−0.36 [−0.63, −0.03]
Response time (RT)		**0.61** [0.39, 0.77]	0.25 [−0.09, 0.54]
Self-ratings			0.38 [0.05, 0.64]
Vision: without references, Attention: random			
Speech intelligibility (SI) score	**−0.71** [−0.84, −0.53]	**−0.87** [−0.92, −0.77]	−0.18 [−0.49, 0.17]
Response time (RT)		**0.71** [0.53, 0.83]	0.09 [−0.25, 0.43]
Self-ratings			0.21 [−0.14, 0.51]
Vision: with references, Attention: fixed			
Speech intelligibility (SI) score	**−0.68** [−0.81, −0.48]	**−0.84** [−0.91, −0.73]	−0.37 [−0.63, −0.04]
Response time (RT)		**0.64** [0.43, 0.79]	0.19 [−0.15, 0.50]
Self-ratings			0.25 [−0.10, 0.54]
Vision: with references, Attention: random			
Speech intelligibility (SI) score	**−0.68** [−0.81, −0.49]	**−0.84** [−0.91, −0.72]	−0.32 [−0.60, 0.03]
Response time (RT)		**0.66** [0.46, 0.80]	0.24 [−0.12, 0.54]
Self-ratings			0.26 [−0.10, 0.55]

Coefficient in bold are significant after Bonferroni correction.

Discussion

Effects of SNR

In this study we examined listening effort during a speech perception task in a group of normal-hearing adults, using three simultaneous measures: verbal RT, task-evoked pupil dilation, and self-ratings. The results show that RTs and self-ratings were both sensitive to the auditory manipulation adopted. As the listening condition became more difficult (i.e. the SNR decreased), participants took significantly longer to repeat the sentence and rated the listening condition as more effortful. These findings are consistent with previous research, in which similar SNRs were investigated (e.g. Prodi and Visentin 2019; Lau et al. 2019).

Pupil dilation analysis revealed a greater overall dilation and a steeper linear slope for lower SNRs. This is also consistent with previous reports indicating that a more cognitively demanding task prompts a greater task-evoked pupil responses (Koelewijn et al. 2014), up to 40–50% SI at least (Wendt et al. 2018). Contrary to our hypothesis, however, task-evoked pupil dilation was not sensitive to all SNRs included in the study. Statistically significant differences were only seen when listening conditions with a SNR of −3 or −6 dB were compared with a SNR of −9 dB. Pupil dilation did not differ statistically between the −3 and −6 dB SNR conditions. In other words, if task-evoked pupil dilation were considered alone, changes in cognitive resource allocation would not appear when the SNR decreases from −3 to −6 dB.

In our paradigm, pupil dilation analysis was limited to the time until the end of the sentence because our concurrent RT measurement produced a response that contaminated the pupil trace after the sentence was completed. This means that some of the information discernible from pupil response (i.e. peak dilation, peak latency) was missing from our analysis, and the portion of the pupil response considered in the analysis represented only a part of the cognitive processing associated with the performance of the listening task. Even though Figure 4 suggests no difference in the pupil dilation trace of the −3 and −6 dB SNR conditions in the time window after the end of the sentence, we cannot exclude that they will have shown no difference had a retention period being included in the paradigm. Similar evidence also emerged in previous studies when pupil dilation was analysed over longer time windows, or its whole time course, including a retention period. When Lau et al. (2019) investigated peak pupil dilation during a sentence recognition task presented in quiet, and at +6 or 0 dB SNR, the mean peak pupil dilation was significantly smaller in quiet than in the two noisy conditions, with no difference between the latter two noisy conditions despite a reduction in task accuracy of almost 50 percentage points. Zekveld et al. (2010) also assessed pupil dilation in a sentence recognition task, with reference to three SI levels (50%, 71%, 84%) obtained by varying the SNR (−4.4, −2.6, and −0.8 dB, approximately). They found pupil dilation greatest in the condition with 50% intelligibility condition, and no difference between the conditions with 71% and 84% intelligibility levels. The Authors argued that a gap of 13 percentage points was not enough to prompt an increase in task-evoked pupil dilation. In the present study, SI increased by almost 15 percentage points between the −3 and −6 dB SNR conditions. Given the sensitivity of pupil response to SI (Zekveld et al. 2010; Wendt et al. 2018), it may be true that a larger difference is needed before any significant changes in task-evoked pupil dilation become apparent, when good-to-excellent SI levels are used.

Relations between RTs, self-ratings and pupil dilation

A second goal of the present study was to examine the relationship between RTs, self-ratings and task-evoked pupil responses, when these measures are obtained simultaneously. Results from the current study suggest pupil dilation did not correlate with other measures of listening effort. This finding support that idea that pupil dilation is neither an objective correlate of self-ratings of listening effort (McGarrigle et al. 2014; Strand et al. 2018; Lau et al. 2019) nor of RTs (McGarrigle et al. 2017a, 2017b), and reinforce the argument that different potential measures of effort tap into different underlying cognitive dimensions (Alhanbali et al. 2019). The lack of correlation between the task-evoked pupil dilation and other measures of effort had already been reported in a work by Strand et al. (2018), in which a self-report measure was found more sensitive to changes in the SNR than physiological or behavioural (dual-task and recall) measures. Likewise, Lau et al. (2019) found that noisy conditions with a 6 dB difference in SNR could be distinguished using self-ratings, but not from peak pupil dilation.

Furthermore, the nonsignificant results of the rmcorr analysis indicated that within-subject changes in pupil dilation yielded by the task demands (i.e. SNR) did not correlate with corresponding changes in the other measures of listening effort. The result aligns with and extends the finding of McGarrigle et al. (2021), showing that changes over time in task-evoked pupil dilation do no associate with changes in self-ratings of effort at a fixed SNR, but are instead more closely related with the subjective experience of tiredness from listening.

Overall, the results of the present study suggest that during the process of effortful listening pupillary response might measure a cognitive dimension different from the one measured by RTs and self-ratings. For instance, it can be speculated that pupil dilation more accurately tracks changes in cognitive resources allocation whereas self-ratings reflect the perceived performance (Herrmann & Johnsrude, 2020).

Effects of cognitive manipulations

Neither of the cognitive manipulations adopted in our task (i.e. focus vs. divided attention; visible vs. invisible references to the location of the loudspeakers) proved effective in influencing our dependent variables. One possible interpretation for this null effect lies in the nature of the masking stimuli used in the ITAMatrix, which were ‘energetic’ rather than ‘informational’ (Pollack 1975). While energetic masking stimuli (such as the stationary noise used in our case) interfere with the speech signal in the acoustic environment and at the hearing periphery (i.e. ‘outside of the perceiver’, as Lidestam et al. 2014 put it), informational masking (such as a concurrent talker) also interferes with the speech signal at cognitive level, in terms of speech information processing (i.e. ‘inside the perceiver, in the perceptual process’). From this perspective, cognitive manipulations may be more effective when using informational rather than energetic masking. A previous study by Koelewijn et al. (2015) examined the effects of attentional instructions (focussed vs. divided) on task-dependent changes in pupil dilation. They found a benefit of focussed attention, in terms of a reduced peak pupil dilation, when they asked participants to repeat a sentence spoken to one ear by a female speaker while ignoring a sentence spoken to the other ear by a male speaker. Although both sentences were masked by independent fluctuating noise, these stimuli combined informational with energetic masking. The Matrix Sentence Test that we adopted, based only on energetic masking, may therefore be suboptimal for the purpose of investigating cognitive manipulations on listening effort. Another aspect that may have influenced the cognitive manipulation was asking participants to direct their gaze towards a central fixation point. While this was a mandatory constraint to have reliable pupillometry measure, it is possible that directing gaze directly towards the visible LEDs signalling the position of the sound source (an overt attention strategy) could have produced measurable effects of our experimental manipulation.

Limitations of the study

Our study findings have potential implications for assessing listening effort using different simultaneous measures, but several factors limit the generalisability of these findings.

First, a fixed level of background noise was used in the experiment, and the acoustic manipulation realised by varying the level of the target speech, so that participants could not anticipate the SNR prior to speech onset. It might be that the opposite choice (varying the level of background noise, with a fixed level of target speech) would result in a pupil trace showing the peak earlier in time, thanks to the anticipatory pupil dilation (McCloy et al. 2017). This would potentially allow for the task-evoked pupil response peak to show during the listening phase (i.e. within the analysis time window), and to be separated from the pupil dilations related to response preparation and delivery.

Second, we used a speech intelligibility task in this study. The effects of auditory and cognitive manipulations on task-related pupil dilation might vary depending on the test material used. For instance, they could be more evident for a task involving a higher level of speech processing (i.e. speech comprehension) or using background noise with an informative content (Wendt et al. 2018).

Third, our study explored the sensitivity of different measures of listening effort under listening conditions that produce different levels of performance. As self-ratings often reflect accuracy rather perceived effort - i.e. participants would rate their perceived performance rather than how much effort they put into a task (Picou and Ricketts 2018) - future works should explore the sensitivity of the three measures under conditions that generate similar performance levels (e.g. adaptive tests with a fixed intelligibility target).

Finally, having participants perform two testing sessions separated by a week might have induced variability due to factors such as the time of the day, participants' mood, or what they have been doing before coming to the experiment. We controlled for this variability by counterbalancing the order of the vision condition across participants, but still in future studies an experimental design limited to a single day could be implemented.

Conclusions

Verbal RT, task-evoked pupil dilation and self-ratings of listening effort were measured simultaneously during a speech intelligibility task, administered using the internationally-validated Matrix Sentence Test. Some limitations to using the simultaneous measures approach were considered and discussed. They mainly refer to the absence of a retention period after the listening phase, so that only a part of the pupillometric response could be reliably included in the analysis.

In the present study, the three measures differed in their sensitivity to experimentally-induced changes in the auditory environment (SNR). RTs and self-ratings proved most sensitive to changes in SNR, thus confirming the utility of a simple—and methodologically less challenging—measure such as verbal RT for assessing listening effort. The three measures used in this study did not show strong relations. In particular, while within-subject changes in RTs yielded by the manipulation of the task demands (i.e. SNR) correlated with the corresponding changes in self-reports, no association was found between either of the two measures and pupillometry changes. The result reinforces the argument that different measures of effort tap into different underlying cognitive dimensions.

Notes

1. To explore if uninformative visual cue could affect auditory attention, participants were tested with and without references about the positions of the sources. The condition without references was obtained by using a curtain occluding the speakers. In order to avoid differences in the stimulus playback with and without the curtain, the curtain was used in both visual conditions and the visual cue provided via LEDs.
2. The R code for the statistical model of speech intelligibility was: _m.intell = glmer(intell ~ SNR*attention*vision+(1|subject)+(SNR + attention + vision|subject),data = data, family = binomial, glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 100000))._
3. The R code for the statistical model of response times was: _mod.RT = glmer(RT ~ SNR*attention*vision+(1 |subject)+(SNR|subject), data = data, family = Gamma(link="log"), glmerControl(optimizer = "bobyqa", optCtrl = list(maxfun = 100000))._
4. The R code for the statistical model of self-ratings was: _mod.ratings = clmm2(rating ~ SNR*attention*vision,random = subject, data = data, Hess = TRUE)_
5. The analysis was also performed with the traces normalized (e.g. within-trial mean scaling, Winn et al. 2018). As no differences emerged in the results of the two analyses, only the analysis with the baseline-adjusted pupil dilation data are presented here.
6. The R code for the statistical model of the pupil dilation was: _m.pupil = lmer(pupil~(ot1)*SNR*attention*vision+(ot1 |subject)+(ot1 |subject:SNR:attention:vision),data = data,control = lmerControl(optimizer="bobyqa"),REML = FALSE)_
7. An analysis performed over a less conservative time window [0; 4 s] returned the same results.

Acknowledgements

The authors acknowledge Hörtech GmbH, Oldenburg, Germany, for providing the speech recordings of the Matrix Sentence Test in the Italian language.

Funding

F.P. and C.Va. are supported by a grant from the Agence Nationale de la Recherche [ANR-16-CE17-0016]. F.P. is supported by grants from the Fondation Neurodis, Fondation MEDISITE and Italian Ministry for the University and Research [PRIN 20177894ZH]. C.Va. is supported by a grant from the Università Italo-Francese (UIF)/Université Franco-Italienne (UFI) and the Ermenegildo Zegna Founder's Scholarship. C.Vi. is supported by the Emilia-Romagna Regional Authority’s _COMPRENDO_ project [Action POR FESR 2014-20 – Axis 1 Action 1.2.2 (2019)].

References

Alhanbali, S., P. Dawes, R. E. Millman, and K. J. Munro. 2019. “Measures of Listening Effort Are Multidimensional.” *Ear and Hearing* 40 (5): 1084–1097. Crossref. PubMed. Web of Science.

Anderson Gosselin, P., and J.-P. Gagné. 2010. “Use of a Dual-Task Paradigm to Measure Listening Effort.” *Revue canadienne d’orthophonie et d’audiologie* 34 (1): 43–51.

Aston-Jones, G., and J. D. Cohen. 2005. “An Integrative Theory of Locus Coeruleus-Norepinephrine Function: Adaptive Gain and Optimal Performance.” *Annual Review of Neuroscience* 28: 403–450. Crossref. PubMed. Web of Science.

Bakdash, J. Z., and L. R. Marusich. 2017. “Repeated Measures Correlation.” *Frontiers in Psychology* 8: 456. Crossref. PubMed. Web of Science.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software* 67 (1): 1–48. Crossref. Web of Science.

Best, V., E. J. Ozmeral, and B. G. Shinn-Cunningham. 2007. “Visually-Guided Attention Enhances Target Identification in a Complex Auditory Scene.” *Journal of the Association for Research in Otolaryngology: JARO* 8 (2): 294–304. Crossref. PubMed.

Boersma, P., and D. Weenink. 2019. *Praat: Doing Phonetics by Computer* [Computer program]. Version 6.0.33, retrieved May 2018 from http://www.praat.org/

Choi, S., A. Lotto, D. Lewis, B. Hoover, and P. Stelmachowicz. 2008. “Attentional Modulation of Word Recognition by Children in a Dual-Task Paradigm.” *Journal of Speech, Language, and Hearing Research* 51 (4): 1042–1054. Crossref. PubMed.

Christensen, R. H. B. 2019. *Ordinal-Regression Models for Ordinal Data*. R package version 2019.4-25, from http://www.cran.r-project.org/package=ordinal/

Driver, J. 2001. “A Selective Review of Selective Attention Research from the past Century.” *British Journal of Psychology* 92 (1): 53–78. Crossref. Web of Science.

Gagné, J.-P., J. Besser, and U. Lemke. 2017. “Behavioral Assessment of Listening Effort Using a Dual-Task Paradigm.” *Trends in Hearing* 21: 2331216516687287. Crossref. Web of Science.

Gatehouse, S., and J. Gordon. 1990. “Response Times to Speech Stimuli as Measures of Benefit from Amplification.” *British Journal of Audiology* 24 (1): 63–68. Crossref. PubMed.

Gustafson, S., R. McCreery, B. Hoover, J. G. Kopun, and P. Stelmachowicz. 2014. “Listening Effort and Perceived Clarity for Normal-Hearing Children with the Use of Digital Noise Reduction.” *Ear and Hearing* 35 (2): 183–194. Crossref. PubMed. Web of Science.

Herrmann, B., and I. S. Johnsrude. 2020. “A Model of Listening Engagement (MoLE).” *Hearing Research* 397: 108016. Crossref. PubMed. Web of Science.

Hornsby, B. W. 2013. “The Effects of Hearing Aid Use on Listening Effort and Mental Fatigue Associated with Sustained Speech Processing Demands.” *Ear and*

Hearing 34 (5): 523–534. Crossref. PubMed. Web of Science.

Hothorn, T., F. Bretz, and P. Westfall. 2008. “Simultaneous Inference in General Parametric Models.” *Biometrical Journal. Biometrische Zeitschrift* 50 (3): 346–363. Crossref. PubMed. Web of Science.

Houben, R., M. van Doorn-Bierman, and W. A. Dreschler. 2013. “Using Response Time to Speech as a Measure for Listening Effort.” *International Journal of Audiology* 52 (11): 753–761. Crossref. PubMed. Web of Science.

Hupé, J. M., C. Lamirel, and J. Lorenceau. 2009. “Pupil Dynamics During Bistable Motion Perception.” *Journal of Vision* 9 (7): 10–10. Crossref. PubMed. Web of Science.

Jaeger, T. F. 2008. “Categorical Data Analysis: Away from ANOVAs (Transformation or Not) and Towards Logit Mixed Models.” *Journal of Memory and Language* 59 (4): 434–446. Crossref. PubMed. Web of Science.

Kahneman, D. 1973. *Attention and Effort*. Englewood Cliffs: Prentice-Hall.

Koelewijn, T., H. de Kluiver, B. G. Shinn-Cunningham, A. A. Zekveld, and S. E. Kramer. 2015. “The Pupil Response Reveals Increased Listening Effort When It is Difficult to Focus Attention.” *Hearing Research* 323: 81–90. Crossref. PubMed. Web of Science.

Koelewijn, T., B. G. Shinn-Cunningham, A. A. Zekveld, and S. E. Kramer. 2014. “The Pupil Response is Sensitive to Divided Attention During Speech Processing.” *Hearing Research* 312: 114–120. Crossref. PubMed. Web of Science.

Koelewijn, T., A. A. Zekveld, J. M. Festen, J. Rönnberg, and S. E. Kramer. 2012. “Processing Load Induced by Informational Masking is Related to Linguistic Abilities.” *International Journal of Otolaryngology* 2012: 865731. Crossref. PubMed.

Koelewijn, T., A. A. Zekveld, T. Lunner, and S. E. Kramer. 2018. “The Effect of Reward on Listening Effort as Reflected by the Pupil Dilation Response.” *Hearing Research* 367: 106–112. Crossref. PubMed. Web of Science.

Kollmeier, B., A. Warzybok, S. Hochmuth, M. A. Zokoll, V. Uslar, T. Brand, and K. C. Wagener. 2015. “The Multilingual Matrix Test: Principles, Applications, and Comparison Across Languages: A Review.” *International Journal of Audiology* 54 (sup2): 3–16. Crossref. PubMed. Web of Science.

Lau, M. K., C. Hicks, T. Kroll, and S. Zupancic. 2019. “Effect of Auditory Task Type on Physiological and Subjective Measures of Listening Effort in Individuals with Normal Hearing.” *Journal of Speech, Language, and Hearing Research : JSLHR* 62 (5): 1549–1560. Crossref. PubMed.

Lemke, U., and J. Besser. 2016. “Cognitive Load and Listening Effort: Concepts and Age-Related Considerations.” *Ear & Hearing* 37 (1): 77S–84S. Crossref. PubMed.

Lenth, L. 2019. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.3.4. https://CRAN.R-project.org/package=emmeans

Lewis, D., K. Schmid, S. O’Leary, J. Spalding, E. Heinrichs-Graham, and R. High. 2016. “Effects of Noise on Speech Recognition and Listening Effort in Children with Normal Hearing and Children with Mild Bilateral or Unilateral Hearing

Loss." *Journal of Speech, Language, and Hearing Research : JSLHR* 59 (5): 1218–1232. Crossref. PubMed. Web of Science.

Lidestam, B., J. Holgersson, and S. Moradi. 2014. "Comparison of Informational vs. Energetic Masking Effects on Speechreading Performance." *Frontiers in Psychology* 5: 639. Crossref. PubMed.

Mattys, S. L., M. H. Davis, A. R. Bradlow, and S. K. Scott. 2012. "Speech Recognition in Adverse Conditions: A Review." *Language and Cognitive Processes* 27 (7-8): 953–978. Crossref. Web of Science.

McCloy, D. R., E. D. Larson, B. Lau, and A. K. Lee. 2016. "Temporal Alignment of Pupillary Response with Stimulus Events via Deconvolution." *The Journal of the Acoustical Society of America* 139 (3): EL57–EL62. Crossref. PubMed.

McCloy, D. R., B. K. Lau, E. Larson, K. A. Pratt, and A. K. Lee. 2017. "Pupillometry Shows the Effort of Auditory Attention Switching." *The Journal of the Acoustical Society of America* 141 (4): 2440–2451. Crossref. PubMed.

McGarrigle, R., P. Dawes, A. J. Stewart, S. E. Kuchinsky, and K. J. Munro. 2017a. "Measuring Listening-Related Effort and Fatigue in School-Aged Children Using Pupillometry." *Journal of Experimental Child Psychology* 161: 95–112. Crossref. PubMed. Web of Science.

McGarrigle, R., P. Dawes, A. J. Stewart, S. E. Kuchinsky, and K. J. Munro. 2017b. "Pupillometry Reveals Changes in Physiological Arousal During a Sustained Listening Task." *Psychophysiology* 54 (2): 193–203. Crossref. PubMed. Web of Science.

McGarrigle, R., S. J. Gustafson, B. W. Hornsby, and F. H. Bess. 2019. "Behavioral Measures of Listening Effort in School-Age Children: Examining the Effects of Signal-to-Noise Ratio, Hearing Loss, and Amplification." *Ear and Hearing* 40 (2): 381–392. Crossref. PubMed. Web of Science.

McGarrigle, R., K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, J. G. Barry, and S. Amitay. 2014. "Listening Effort and Fatigue: what Exactly Are We Measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'." *International Journal of Audiology* 53 (7): 433–440. Crossref. PubMed. Web of Science.

McGarrigle, R., L. Rakusen, and S. Mattys. 2021. "Effortful Listening Under the Microscope: Examining Relations Between Pupillometric and Subjective Markers of Effort and Tiredness from Listening." *Psychophysiology* 58 (1): e13703. Crossref. PubMed.

Mealings, K. T., K. Demuth, J. M. Buchholz, and H. Dillon. 2015. "The Effect of Different Open Plan and Enclosed Classroom Acoustic Conditions on Speech Perception in Kindergarten Children ." *The Journal of the Acoustical Society of America* 138 (4): 2458–2469. Crossref. PubMed. Web of Science.

Meister, H., S. Rählmann, U. Lemke, and J. Besser. 2018. "Verbal Response Times as a Potential Indicator of Cognitive Load during Conventional Speech Audiometry with Matrix Sentences." *Trends in Hearing* 22: 2331216518793255. Crossref.

Mirman, D. 2014. *Growth Curve Analysis and Visualization Using R*. Boca Raton: CRC Press.

Ohlenforst, B., D. Wendt, S. E. Kramer, G. Naylor, A. A. Zekveld, and T. Lunner. 2018. “Impact of SNR, Masker Type and Noise Reduction Processing on Sentence Recognition Performance and Listening Effort as Indicated by the Pupil Dilation Response.” *Hearing Research* 365: 90–99. Crossref. PubMed. Web of Science.

Pals, C., A. Sarampalis, H. van Rijn, and D. Başkent. 2015. “Validation of a Simple Response-Time Measure of Listening Effort.” *The Journal of the Acoustical Society of America* 138 (3): EL187–EL192. Crossref. PubMed. Web of Science.

Peelle, J. E. 2018. “Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior.” *Ear and Hearing* 39 (2): 204–214. Crossref. PubMed. Web of Science.

Pichora-Fuller, M. Kathleen, Sophia E. Kramer, Mark A. Eckert, Brent Edwards, Benjamin W. Y. Hornsby, Larry E. Humes, Ulrike Lemke, et al. 2016. “Hearing Impairment and Cognitive Energy: The Framework for Understanding Effortful Listening (FUEL).” *Ear & Hearing* 37 (1): 5S–27S. Crossref. PubMed.

Picou, E. M., and T. A. Ricketts. 2014. “The Effect of Changing the Secondary Task in Dual-Task Paradigms for Measuring Listening Effort.” *Ear and Hearing* 35 (6): 611–622. Crossref. PubMed. Web of Science.

Picou, E. M., and T. A. Ricketts. 2018. “The Relationship Between Speech Recognition, Behavioural Listening Effort, and Subjective Ratings.” *International Journal of Audiology* 57 (6): 457–467. Crossref. PubMed. Web of Science.

Piquado, T., D. Isaacowitz, and A. Wingfield. 2010. “Pupillometry as a Measure of Cognitive Effort in Younger and Older Adults.” *Psychophysiology* 47 (3): 560–569. Crossref. PubMed. Web of Science.

Pollack, I. 1975. “Auditory Informational Masking.” *The Journal of the Acoustical Society of America* 57 (S1): S5–S5. Crossref.

Prodi, N., and C. Visentin. 2019. “Impact of Background Noise Fluctuation and Reverberation on Response Time in a Speech Reception Task.” *Journal of Speech, Language, and Hearing Research : JSLHR* 62 (11): 4179–4195. Crossref. PubMed.

Prodi, N., C. Visentin, E. Borella, I. C. Mammarella, and A. Di Domenico. 2019a. “Noise, Age and Gender Effects on Speech Intelligibility and Sentence Comprehension for 11- to 13-Year-Old Children in Real Classrooms.” *Frontiers in Psychology* 10: 2166. Crossref. PubMed.

Prodi, N., C. Visentin, A. Peretti, J. Griguolo, and G. B. Bartolucci. 2019b. “Investigating Listening Effort in Classrooms for 5- to 7-Year-Old Children.” *Language, Speech, and Hearing Services in Schools* 50 (2): 196–210. Crossref. PubMed. Web of Science.

Puglisi, G. E., A. Warzybok, S. Hochmuth, C. Visentin, A. Astolfi, N. Prodi, and B. Kollmeier. 2015. “An Italian Matrix Sentence Test for the Evaluation of Speech Intelligibility in Noise.” *International Journal of Audiology* 54 (sup2): 44–50. Crossref. PubMed. Web of Science.

R Core Team. 2013. R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/

Steel, M. M., B. C. Papsin, and K. A. Gordon. 2015. “Binaural Fusion and Listening Effort in Children Who Use Bilateral Cochlear Implants: A Psychoacoustic and Pupillometric Study.” *PLoS One* 10 (2): e0117611. Crossref. PubMed. Web of Science.

Strand, J. F., V. A. Brown, M. B. Merchant, H. E. Brown, and J. Smith. 2018. “Measuring Listening Effort: Convergent Validity, Sensitivity, and Links with Cognitive and Personality Measures.” *Journal of Speech, Language, and Hearing Research: JSLHR* 61 (6): 1463–1486. Crossref. PubMed. Web of Science.

Uslar, V. N., R. Carroll, M. Hanke, C. Hamann, E. Ruigendijk, T. Brand, and B. Kollmeier. 2013. “Development and Evaluation of a Linguistically and Audiologically Controlled Sentence Intelligibility Test*.” The Journal of the Acoustical Society of America* 134 (4): 3039–3056. Crossref. PubMed.

van den Tillaart-Haverkate, M., I. de Ronde-Brons, W. A. Dreschler, and R. Houben. 2017. “The Influence of Noise Reduction on Speech Intelligibility, Response Times to Speech, and Perceived Listening Effort in Normal-Hearing Listeners.” *Trends in Hearing* 21: 2331216517716844. Crossref.

van Rij, J., P. Hendriks, H. van Rijn, R. H. Baayen, and S. N. Wood. 2019. “Analyzing the Time Course of Pupillometric Data.” *Trends in Hearing* 23: 2331216519832483. Crossref. Web of Science.

Visentin, C., N. Prodi, F. Cappelletti, S. Torresin, and A. Gasparella. 2018. “Using Listening Effort Assessment in the Acoustical Design of Rooms for Speech.” *Building and Environment* 136: 38–53. Crossref.

Visentin, C., N. Prodi, F. Cappelletti, S. Torresin, and A. Gasparella. 2019. “Speech Intelligibility and Listening Effort in University Classrooms for Native and Non-Native Italian Listeners.” *Building Acoustics* 26 (4): 275–291. Crossref.

Wagner, A. E., L. Nagels, P. Toffanin, J. M. Opie, and D. Başkent. 2019. “Individual Variations in Effort: Assessing Pupillometry for the Hearing Impaired.” *Trends in Hearing* 23: 2331216519845596. Crossref.

Wendt, D., T. Dau, and J. Hjortkjaer. 2016. “Impact of Background Noise and Sentence Complexity on Processing Demands during Sentence Comprehension.” *Frontiers in Psychology* 7: 345. Crossref. PubMed. Web of Science.

Wendt, D., T. Koelewijn, P. Książek, S. E. Kramer, and T. Lunner. 2018. “Toward a More Comprehensive Understanding of the Impact of Masker Type and Signal-to-Noise Ratio on the Pupillary Response While Performing a Speech-In-Noise Test .” *Hearing Research* 369: 67–78. Crossref. PubMed. Web of Science.

Whelan, R. 2008. “Effective Analysis of Reaction Time Data.” *The Psychological Record* 58 (3): 475–482. Crossref. Web of Science.

Winn, M. B. 2016. “Rapid Release from Listening Effort Resulting from Semantic Context, and Effects of Spectral Degradation and Cochlear Implants.” *Trends in Hearing* 20: 233121651666972. Crossref. Web of Science.

Winn, M. B., J. R. Edwards, and R. Y. Litovsky. 2015. “The Impact of Auditory Spectral Resolution on Listening Effort Revealed by Pupil Dilation.” *Ear and Hearing* 36 (4): e153–e165. Crossref. PubMed. Web of Science.

Winn, M. B., D. Wendt, T. Koelewijn, and S. E. Kuchinsky. 2018. “Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started.” *Trends in Hearing* 22: 2331216518800869. Crossref. Web of Science.

Zekveld, A. A., T. Koelewijn, and S. E. Kramer. 2018. “The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge.” *Trends in Hearing* 22: 2331216518777174. Crossref. Web of Science.

Zekveld, A. A., S. E. Kramer, and J. M. Festen. 2010. “Pupil Response as an Indication of Effortful Listening: The Influence of Sentence Intelligibility.” *Ear and Hearing* 31 (4): 480–490. Crossref. PubMed. Web of Science.

Zekveld, A. A., M. Rudner, S. E. Kramer, J. Lyzenga, and J. Rönnberg. 2014. “Cognitive Processing Load during Listening is Reduced More by Decreasing Voice Similarity than by Increasing Spatial Separation Between Target and Masker Speech.” *Frontiers in Neuroscience* 8: 88. Crossref. PubMed. Web of Science.