The Effects of Task Difficulty Predictability and Noise Reduction on Recall Performance and Pupil Dilation Responses

Andreea Micula,[1,2] Jerker Rönnberg,[2] Lorenz Fiedler,[3] Dorothea Wendt,[3,4] Maria Cecilie Jørgensen,[5] Ditte Katrine Larsen,[5] and Elaine Hoi Ning Ng[1,2]

Objectives: Communication requires cognitive processes which are not captured by traditional speech understanding tests. Under challenging listening situations, more working memory resources are needed to process speech, leaving fewer resources available for storage. The aim of the current study was to investigate the effect of task difficulty predictability, that is, knowing versus not knowing task difficulty in advance, and the effect of noise reduction on working memory resource allocation to processing and storage of speech heard in background noise. For this purpose, an "offline" behavioral measure, the Sentence-Final Word Identification and Recall (SWIR) test, and an "online" physiological measure, pupillometry, were combined. Moreover, the outcomes of the two measures were compared to investigate whether they reflect the same processes related to resource allocation.

Design: Twenty-four experienced hearing aid users with moderate to moderately severe hearing loss participated in this study. The SWIR test and pupillometry were measured simultaneously with noise reduction in the test hearing aids activated and deactivated in a background noise composed of four-talker babble. The task of the SWIR test is to listen to lists of sentences, repeat the last word immediately after each sentence and recall the repeated words when the list is finished. The sentence baseline dilation, which is defined as the mean pupil dilation before each sentence, and task-evoked peak pupil dilation (PPD) were analyzed over the course of the lists. The task difficulty predictability was manipulated by including lists of three, five, and seven sentences. The test was conducted over two sessions, one during which the participants were informed about list length before each list (predictable task difficulty) and one during which they were not (unpredictable task difficulty).

Results: The sentence baseline dilation was higher when task difficulty was unpredictable compared to predictable, except at the start of the list, where there was no difference. The PPD tended to be higher at the beginning of the list, this pattern being more prominent when task difficulty was unpredictable. Recall performance was better and sentence baseline dilation was higher when noise reduction was on, especially toward the end of longer lists. There was no effect of noise reduction on PPD.

Conclusions: Task difficulty predictability did not have an effect on resource allocation, since recall performance was similar independently of whether task difficulty was predictable or unpredictable. The higher sentence baseline dilation when task difficulty was unpredictable likely reflected a difference in the recall strategy or higher degree of task engagement/alertness or arousal. Hence, pupillometry captured processes which the SWIR test does not capture. Noise reduction frees up resources to be used for storage of speech, which was reflected in the better recall performance and larger sentence baseline dilation toward the end of the list when noise reduction was on. Thus, both measures captured different temporal aspects of the same processes related to resource allocation with noise reduction on and off.

Key words: Free recall, Noise reduction, Pupillometry, Task difficulty predictability, Working memory.

(Ear & Hearing 2021;42;1668–1679)

INTRODUCTION

Hearing aid users often report that even when speech is loud enough to be understood, listening is effortful (Pichora-Fuller et al. 2016). However, communication involves multiple processes besides listening and understanding speech. Resources also need to be allocated for preparing a response or storing speech in memory, for instance. Thus, communication involves cognitive processes such as working memory (WM) and attention (Kahneman 1973; Koelewijn et al. 2012; Rönnberg et al. 2013; Ng et al. 2013, 2015; Pichora-Fuller et al. 2016).

Traditional speech recognition tests do not reflect the allocation of resources to multiple cognitive processes. Resource allocation during listening tasks has been investigated using various measures, such as behavioral measures (e.g., recall performance) or physiological measures (e.g., pupillometry) (Zekveld et al. 2011; Pichora-Fuller et al. 2016; Wendt et al. 2017; Ohlenforst et al. 2018; Peelle 2018; Zhang et al. 2021). Resource allocation can be modulated by factors such as hearing status, background noise, hearing aid signal processing and task demands (e.g., number of to be recalled words) (Kahneman 1973; Gosselin & Gagné 2011; Zekveld et al. 2011; Lemke & Besser 2016; Pichora-Fuller et al. 2016; Zekveld et al. 2018a). The Ease of Language Understanding (ELU) model (Rönnberg et al. 2008, 2013, 2019) describes the allocation of WM resources during communication.

Resource Allocation According to the ELU Model

There is an increasing amount of evidence supporting the relationship between speech recognition and relevant cognitive abilities, such as WM (Arehart et al. 2015; Gordon-Salant & Cole 2016; Dryden et al. 2017). WM is the ability to simultaneously process and store information (Baddeley 2012). The ELU model describes the role of WM in a communicative context. According to the ELU model, speech input enters an episodic

[1]Oticon A/S, Smørum, Denmark; [2]Department of Behavioural Sciences and Learning, Linnaeus Centre HEAD, Swedish Institute for Disability Research, Linköping University, Linköping, Sweden; [3]Eriksholm Research Centre, Snekkersten, Denmark; [4]Hearing Systems, Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, Kongens Lyngby, Denmark; and [5]Department of Nordic Studies and Linguistics, University of Copenhagen, Amager, Denmark.

<zdoi; 10.1097/AUD.0000000000001053>

buffer, which is referred to as the Rapid, Automatic, Multimodal Binding of PHOnology (RAMBPHO). Under optimal listening conditions, processing is implicit, since RAMBPHO rapidly and automatically matches the speech input to phonological representations stored in semantic long-term memory, which gives access to the mental lexicon. In this case, the amount of WM resources required to process speech is low. When the speech input is degraded due to factors such as background noise or hearing impairment, a mismatch may occur, requiring explicit processing to remediate it. Explicit processing draws on additional processing resources by retrieving information stored in semantic and/or episodic long-term memory to resolve the mismatch (Rönnberg et al. 2008, 2013; Wingfield et al. 2015; Edwards 2016; Rönnberg et al. 2019). Since the WM system is limited in capacity, the resources are allocated to processing and storage within that capacity. Consequently, when more resources are needed to process heard speech, less resources are available for storage (Lunner et al. 2009).

Several measures have been used to capture the effects of adverse listening conditions and task demands on resource allocation (Peelle 2018). In the current study, we combined the Sentence-Final Word Identification and Recall (SWIR) test (Ng et al. 2013, 2015; Micula et al. 2020) with pupillometry (Wendt et al. 2017; Ohlenforst et al. 2018; Zhang et al. 2021; Bönitz et al. 2021), which are described individually in the following sections.

The SWIR Test as a Measure of Resource Allocation

The SWIR test has been developed to measure the effects of hearing aid signal processing on recall of intelligible speech in background noise (Ng et al. 2013, 2015). The task of the SWIR test is to listen to lists of sentences, repeat the last word of each sentence immediately after it is finished and recall as many of the repeated words as possible when the list is finished. The test is administered at 95% speech intelligibility. The first studies using the SWIR test (Ng et al. 2013, 2015) showed that recall performance in competing speech was better when noise reduction was activated compared to when it was not. Noise reduction presumably facilitates segregation of target speech from background noise and thus facilitates implicit processing, which leads to a decrease in the amount of WM resources needed for processing speech, allowing additional resources to be used for storage. Thus, the SWIR test can be seen as a measure of allocation of WM resources to storage of heard speech under various processing demands.

A recent study (Micula et al. 2020) has replicated the findings by Ng et al. on the effects of background noise and noise reduction on resource allocation. In addition, this study investigated the effects of task difficulty by varying the list length and task difficulty predictability. Micula et al. (2020) included a group of participants who were informed about the length of the upcoming list (predictable task difficulty) and a group of participants who were not informed about list length in advance (unpredictable task difficulty), although the participants knew that it could vary. The authors considered unpredictable task difficulty to be more ecologically valid because utterance length is essentially unknown in real life dialog. Since this was the first study manipulating the task difficulty of the SWIR test, understanding the effect of task difficulty predictability was relevant for assessing the ecological validity of the outcomes. The findings showed that the SWIR test with varying task difficulty, that is, list length, was sensitive to detecting the benefit of noise reduction. On the other hand, task difficulty predictability, that is, knowledge about list length, did not have an effect on speech intelligibility and overall recall performance, indicating that task difficulty predictability does not have an effect on WM resource allocation to processing and storage of speech. However, listeners tended to recall the words in a different order when noise reduction was off and task difficulty was unpredictable. Although knowledge about list length does not seem to influence overall recall performance, participants cannot choose a strategy based on task difficulty when list length is unknown (Grenfell-Essam & Ward 2012). This suggests that task difficulty predictability may have an effect on cognitive processes other than resource allocation, such as the strategy used to encode words (Micula et al. 2020).

Pupillometry as a Measure of Resource Allocation

It is well-established that the magnitude of pupil dilation reflects the mental effort or processing resources needed to perform a certain cognitive task (Beatty 1982). Pupil dilation is an index of the noradrenergic function of the locus coeruleus, which is associated with physiological responses to arousal, stress, emotions, as well as to various cognitive functions and optimization of behavioral task performance. The neurons in the locus coeruleus have a tonic mode, which reflects baseline activity, and a phasic mode, which reflects activity in response to a stimulus. Thus, baseline pupil dilation is associated with the tonic mode, while task-evoked pupil dilation is associated with the phasic mode (Aston-Jones & Cohen 2005; Laeng et al. 2012; Winn et al. 2018; Zekveld et al. 2018a). Pupillary responses occur spontaneously and involuntarily (Laeng et al. 2012). Therefore, pupillometry has become a widely used objective measure of resource allocation during speech recognition tasks. Moreover, pupillometry is a time-series measurement, thus making it possible to follow changes in resource allocation over the time course of a task (Winn et al. 2018). Unsworth and Robison (2014) refer to pupillary responses as an "online" measure of attentional resource allocation. Hence, we are using the term "online" in this article to refer the method of tracking changes in pupillary responses over the course of the SWIR test list. In comparison, the SWIR test is rather an "offline" measure, since it yields an overall score of behavioral performance.

Previous research has demonstrated that noise reduction in hearing aids leads to a decrease in task-evoked pupil dilation during a speech recognition task administered in competing speech, even at very high speech intelligibility levels (Wendt et al. 2017; Ohlenforst et al. 2018). In these studies, the peak pupil dilation (PPD) was analyzed, which is a phasic pupillary response measured as the maximum pupil dilation triggered by the task stimulus. According to Wendt et al. (2017) and Ohlenforst et al. (2018), a decrease in PPD presumably reflects a decrease in allocation of resources to speech processing. Consequently, these studies show that even when speech intelligibility is at ceiling and noise reduction cannot contribute to better speech recognition, noise reduction can still provide a cognitive benefit. In terms of the ELU model, the lower PPD when noise reduction is activated can be interpreted as a facilitation of implicit processing, since less WM resources are needed to process speech (Rönnberg et al. 2013; Wendt et al., 2017; Rönnberg et al. 2019). Interestingly, some studies have demonstrated that individuals with higher WM capacity exhibit greater pupil dilation as well as better speech recognition

performance in adverse listening conditions than individuals with lower WM capacity (Zekveld et al. 2011; Koelewijn et al. 2012). Increased pupil dilation is believed to be related to mobilization of a higher amount of WM resources or explicit processing. This can be interpreted as a more efficient allocation of WM resources, such that people with high WM capacity are able to exert a higher degree of task engagement to overcome challenges encountered in difficult listening conditions. In sum, pupil dilation responses seem to be affected by various mechanisms of resource allocation (Rönnberg et al. 2013, 2019).

Two studies that have combined speech recognition and recall tasks with pupillometry to investigate online resource allocation are reviewed in the following sections. In their recent study, Zhang et al. (2021) administered a word recognition and recall task in various degrees of background noise to a group of young participants with normal hearing. The participants were informed before each list whether recall would be required or not. Both baseline dilation and PPD relative to baseline were analyzed. The findings showed that baseline dilation decreased over the course of the list when only repetition was required, but it increased when words also had to be recalled. On the other hand, the PPD decreased for both task conditions, the decrease being steeper when recall was required compared to when it was not. The authors interpreted the increasing baseline dilation as an index of the amount of resources that were allocated to the additional task, that is, maintaining the words in memory for subsequent recall. The decrease in PPD, however, which seems to contradict the well-established pattern of resource allocation to speech recognition, was seen as evidence that the more resources were required for maintaining words in memory throughout the list, the fewer resources were available for speech understanding.

Bönitz et al. (2021) have conducted a study combining pupillometry and a version of the SWIR test in German with lists of three and six sentences in a background noise of four-talker (4T) babble at high speech intelligibility levels. They included older participants with normal hearing. In their study, the sentence baseline dilation and task-evoked mean pupil dilation relative to sentence baseline were analyzed. Moreover, the authors investigated the effects of noise reduction on recall performance, sentence baseline dilation, and mean pupil dilation. They did not find any significant effect of noise reduction on recall performance and sentence baseline dilation. However, mean pupil dilation was overall significantly lower when noise reduction was on compared to off, which the authors interpreted as a decrease in resources required to process speech. This is in line with previous findings on the effect of noise reduction on resource allocation (Wendt et al. 2017; Ohlenforst et al. 2018). Additionally, the findings showed that the sentence baseline decreased over the course of the list when words did not have to be recalled. However, the sentence baseline increased over the course of the list when words had to be recalled and this increase was steeper for longer lists compared to shorter lists. These findings are in line with those by Zhang et al. (2021) and provide further evidence for the sentence baseline as an index of resource allocation to storage of speech.

The Aims of the Current Study

The current study was designed to build on the findings by Micula et al. (2020) by complementing the SWIR test outcomes with pupillometry outcomes. The findings from studies using the SWIR test and studies combining pupillometry with speech recognition tasks highlight the importance of measuring resource allocation in addition to speech intelligibility performance, as some of the potential benefits of hearing aid signal processing cannot be captured by the latter measure.

The first aim was to investigate the effect of task difficulty predictability on recall performance and pupil dilation. For this purpose, the task difficulty of SWIR test was varied so that the lists contained three, five, or seven sentences. Despite not finding a significant effect of task difficulty predictability on overall recall performance, the outcomes from the study by Micula et al. (2020) showed that words were recalled in a different order when task difficulty was unpredictable and noise reduction was off. This suggests that task difficulty predictability may have had an effect on other cognitive processes, such as recall strategy, rather than on WM resource allocation. Based on the study by Micula et al. (2020), we do not expect to find an effect of task difficulty predictability on overall recall performance. To our knowledge, no similar studies have investigated the effect of task difficulty predictability on pupil dilation in a group of participants with hearing impairment using a recall task. Based on the behavioral findings by Micula et al. (2020) and assuming that the PPD reflects the amount of resources used for speech processing and baseline dilation reflects the amount of resources allocated for storage (Wendt et al. 2017; Ohlenforst et al. 2018; Zhang et al. 2021; Bönitz et al. 2021), we do not expect to find an effect of task difficulty predictability on these pupil indices. We hypothesize that any changes in sentence baseline dilation or PPD reflect cognitive processes other than WM resource allocation, since different processes can result in similar pupillary responses (Kahneman 1973).

The second aim was to investigate the effect of noise reduction on recall performance and pupil dilation. Although the effect of noise reduction on recall has been investigated in various studies (Ng et al. 2013, 2015; Neher et al. 2018; Micula et al. 2020), the addition of pupillometry is novel. The present study was conducted in a background noise composed of 4T babble, based on evidence that steady state noise is less cognitively demanding and therefore does not affect resource allocation to the extent to which competing speech does (Ng et al., 2013 2015; Ohlenforst et al. 2018; Micula et al. 2020). Based on previous research using the SWIR test in groups of hearing aid users, we hypothesize that recall performance will be better when noise reduction is activated compared to when it is not (Ng et al. 2013, 2015; Micula et al. 2020). Although Bönitz et al. (2021) did not find an effect of noise reduction on recall performance using the SWIR test, this may have been due to the participants having normal hearing. Some studies have shown that recall performance is influenced by hearing status, such that older adults with normal audiometric thresholds perform significantly better on recall tasks than older adults with hearing loss (Smith et al. 2016), even when there are no differences in speech recognition (McCoy et al. 2005). Regarding pupillary responses, we expect findings similar to those of previous studies combining pupillometry with a speech recognition task (Wendt et al. 2017; Ohlenforst et al. 2018). We hypothesize that the PPD will be smaller when noise reduction is activated compared to when it is not, based on evidence showing that noise reduction decreases the amount of resources needed for speech processing (Ng et al. 2013, 2015; Lunner et al. 2016; Micula et al. 2020). Furthermore, based on recent studies providing

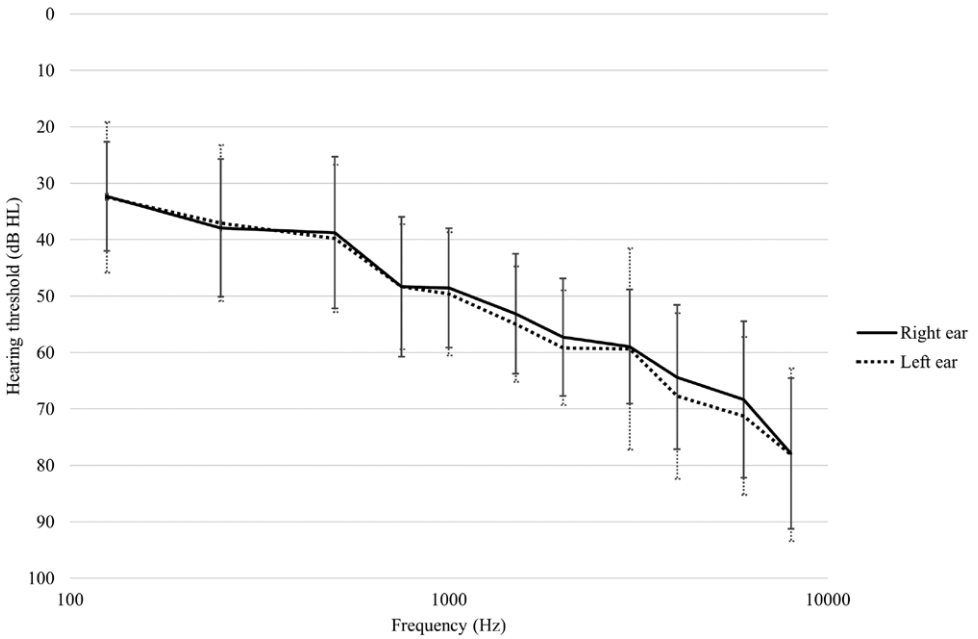

Fig. 1. Mean audiometric thresholds across 125 to 8000 Hz for the right (continuous line) and left (dotted line) ears. The error bars show the standard deviation.

evidence for baseline pupil dilation as an index of resource allocation to maintenance of speech in memory (Zhang et al. 2021; Bönitz et al. 2021), we hypothesize that the sentence baseline dilation will be higher when noise reduction is on, due to the possibility to allocate more resources to the recall task.

Since pupillometry is an "online" measure of resource allocation, while the SWIR test is an "offline" measure of resource allocation, they may capture different temporal aspects of the same cognitive processes involved in processing and storing speech. Therefore, the third aim of this study was to investigate whether two different but interrelated measures, the SWIR test and pupillometry, capture the same cognitive processes related to resource allocation. We speculate that if the effects of task difficulty predictability or noise reduction are captured by both behavioral performance and pupillary responses, then the SWIR test and pupillometry provide "offline" and "online" information respectively of the same processes related to resource allocation. Although Bönitz et al. (2021) have conducted a study combining the SWIR test and pupillometry to investigate resource allocation during speech understanding and recall in individuals with normal hearing, no previous studies have done this including individuals with hearing impairment. We hypothesize that the two measures will provide different temporal information of the same processes when it comes to the effect of noise reduction on resource allocation (Ng et al. 2013, 2015; Wendt et al. 2017; Ohlenforst et al. 2018; Micula et al. 2020; Bönitz et al. 2021). Regarding the effect of task difficulty predictability, it is unknown whether the two measures will be in agreement, as no studies have investigated this combining the SWIR test and pupillometry before.

MATERIALS AND METHODS

Participants

Twenty-four native Danish speakers (11 females and 13 males) with moderate to moderately severe symmetrical sensorineural hearing loss were recruited for participation in this study from the database at Oticon A/S, Smørum, Denmark. Their mean age was 65 years (SD = 8.00, range: 43–75 years) and their average pure-tone thresholds (PTA) at 0.5, 1, 2, and 4 kHz were 52.24 dB HL (SD = 8.03) on the right ear and 54.06 dB HL (SD = 8.26) on the left ear (Fig. 1). All participants were experienced hearing aids users. Their vision was normal or corrected to normal and they did not have a history of eye disease or eye surgery. The Research Ethics Committees of the Capital Region of Denmark have assessed that the current study is exempt from application for ethical approval. All participants signed a written consent form, and the study was conducted according to the standards set by the Declaration of Helsinki.

Assessment Tools

The SWIR Test • The SWIR test was developed by Ng et al. (2013, 2015) to measure the effect of noise reduction on memory for highly intelligible speech in background noise. The task consists of listening to lists of sentences, repeating the last word immediately after each sentence and at the end of the list, indicated by a beep tone, recalling all of the repeated words. The test yields an identification score (percentage of repeated words) and a recall score (percentage of correctly recalled words). It should be noted that misheard words were accepted if they were correctly recalled. In the current study, a Danish version of the SWIR test was used, which is composed of Danish HINT sentences (Nielsen & Dau 2011). Similarly to a previous implementation of the Swedish SWIR test with varying list length (Micula et al. 2020), lists of three, five, and seven sentences were included in the current study.

Pupillometry • The pupil dilation responses were recorded using the iView X RED (Senso-Motoric Instruments) eye-tracker, which tracks both eye and head movement without contact via an infrared camera. The sampling frequency was 60 Hz. The illumination was set to 250 lux. Changes in pupil dilation during each sentence for each participant were recorded for the right and left eyes. If 50% of the data for a sentence was

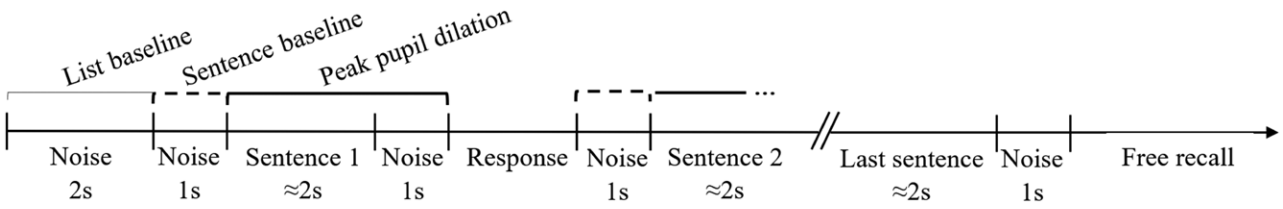

Fig. 2. Example of the time course of a list including the time windows during which list baseline dilation (continuous thin solid line), sentence baseline dilation (dotted line), and peak pupil dilation (continuous thick solid line) were calculated.

missing, likely due to blinks or head movements, the respective sentence was discarded (deblinking). If after deblinking the ratio of missing data for one eye exceeded 40%, all data for the respective eye was excluded. The pupil dilation responses of every participant were averaged across sentences for each condition. Moreover, the pupil dilation responses were averaged for the right and left eyes when data for both eyes was available. The sentence baseline dilation and PPD were calculated for each condition. The sentence baseline dilation was defined as the mean pupil dilation during the one second of noise before the beginning of the sentence. The PPD was considered to be the maximum pupil dilation that occurred between the start of the sentence presentation and the end of the one second of noise after the sentence ended. The PPD was corrected for sentence baseline dilation. Furthermore, both sentence baseline dilation and PPD were corrected for the list baseline, which is the mean pupil dilation of the first two seconds of noise (out of three seconds) before the first sentence of a list. Figure 2 illustrates the timing of these pupil indices over the course of a list.

Test Conditions

The SWIR test and pupillometry were conducted in six test conditions: three list lengths and two noise reduction conditions. These conditions were repeated during two test sessions that were almost identically structured. During one session, the participants were told in advance how many sentences they would hear in the following list (predictable task difficulty), while during the other they were not informed about list length (unpredictable task difficulty). The order of the test sessions was randomized for each participant. The two sessions were scheduled at least three weeks apart to avoid learning effects of the HINT sentences (Nielsen & Dau 2011).

SWIR Test List Length • Micula et al. (2020) have previously conducted a study implementing a version of the SWIR test with varying task difficulty in Swedish. In the current study using the Danish SWIR test, lists of three, five, and seven sentences were included, which allows to manipulate task difficulty predictability. The lists of varying length were obtained by rearranging the sentences of the SWIR test version with seven-sentence lists (Lunner et al. 2016) into new lists. The lists were constructed so that the frequency of the last word was on average similar across lists. All sentences were repeated twice (105 × 2), but in list combinations of different sentences and list lengths. Since the SWIR test was administered in six conditions and seven list repetitions were included per condition, 42 lists were formed. In addition, five lists of seven sentences were administered for procedural training. The sentence material is composed of recordings of a male speaker. The Danish SWIR test shares the same sentence corpus as the Danish version of the HINT (Nielsen & Dau 2011). However, none of the sentences used in the HINT in the current study were included in the SWIR test.

Hearing Aids and Noise Reduction • For the test, all participants were fitted with commercially available hearing aids (Oticon OpnS1 TM mini-Receiver-in-the-ear) and power domes. The hearing aids were fitted using proprietary software based on each participant's PTA thresholds before the test session. In the conditions where noise reduction was off, only amplification was provided using the Voice Aligned Compression (VAC+) rationale, which is a quasilinear fitting rationale based on the loudness data from Buus & Florentine (2001). The VAC+ rationale has lower compression kneepoints to provide more compression at low input levels and less compression at high input levels. A microphone setting that is close to omni-directional was chosen to simulate the acoustic effect of the pinna. In the conditions when noise reduction was on, a fast-acting version of a minimum-variance distortionless response beam-former was applied, which uses spatial filtering in order to attenuate background noise coming from behind the listener (Kjems & Jensen 2012). Additionally, a single-channel Wiener postfilter was applied to further attenuate background noise (Jensen & Pedersen 2015). Wendt et al. (2017) and Ohlenforst et al. (2018) have likewise used these settings.

Background Noise • Previous research using the SWIR test has demonstrated that the effect of noise reduction on recall is most effective in competing speech (Ng et al. 2013, 2015; Micula et al. 2020). Therefore, 4T babble was used in the current study, which was composed of recordings of two male and two female native Danish speakers reading different passages of a newspaper article. The long-term average spectra of the 4T babble resemble that of the HINT sentences. The background noise started three seconds before the first sentence in the SWIR test list and one second before the remaining sentences in the list. The background noise stopped one second after the sentence ended.

Test Setup

The test participants were seated in a sound-proof anechoic chamber. The setup was similar to the one in the study by Wendt et al. (2017). The target speech was presented at a level of 65 dB SPL from a loudspeaker placed at 0° and the 4T babble at an individualized level was presented from four loudspeakers placed at ±90° and ±150°. All loudspeakers were placed at a distance of 1.2 m from the test participant. The eye-tracker was placed in between the front loudspeaker and the participant at a distance individually adjusted based on optimal pupil detection. The hearing aids were connected to a test computer via a programming device. The test computer was outside of the chamber, hence allowing the tester to control the noise reduction settings.

Procedure

At the beginning of the first session, the HINT was administered to establish the SNR corresponding to 80% speech

intelligibility. This was done by using a modified procedure, so that the SNR was decreased by 0.8 dB after correct repetition and increased by 3.2 dB after incorrect repetition, except for the first five sentences, for which the step size was twice as large. Both target speech and 4T babble were at a level of 65 dB SPL at the beginning of the test. Throughout the HINT the target speech remained constant and the 4T babble fluctuated based on the participants' responses. This method has been used by Micula et al. (2020) to approximately estimate the SNR level to be used in the SWIR test. To achieve 95% word recognition in the SWIR test, four training lists were administered for procedural training, as well as for further individual adjustment of the SNR level if needed. This is done by leaving the SNR obtained from the HINT unchanged if six or seven (86–100%) last words from the SWIR test training list are repeated correctly, increasing the SNR by 1 dB if four or five words are repeated correctly or increasing the SNR by 2 dB if zero to three words are repeated correctly (Lunner et al. 2016; Micula et al. 2020). A fifth training list was administered after introducing the instructions for the pupillometry measure. The SWIR test training was administered with noise reduction off. The SNR obtained after the fourth training list was used for the remainder of the SWIR test during the first test session. The mean SNR for all test participants at which the SWIR test was conducted was 6.95 dB (SD = 5.76). The HINT and SWIR test training were repeated at the beginning of the second session, for the participants to get equal amount of training. However, the SWIR test during the second session was administered at the same SNR as in the first session. During the SWIR test, the two noise reduction conditions and the list length were randomized.

Statistical Analysis

An analyses of variance (ANOVA) was performed with noise reduction, task difficulty predictability, and list length as within-subject factors to investigate their effect on identification performance. Another ANOVA was conducted with the same within-subject factors in order to investigate their effects on recall performance.

Since pupillometry is a time-series measurement, the change in pupil dilation over the course of a list was of interest. Bönnitz et al. (accepted) used a linear fit to calculate slope coefficients for this purpose. However, since the authors showed that their data was not linear, a method similar to the one used by Zhang et al. (2021) was chosen for the current study. Thus, the sentence baseline dilation and PPD were analyzed per sentence in the list (serial position). Three sets of ANOVAs were performed for sentence baseline dilation, one for each list length, with task difficulty predictability, noise reduction and serial position as within-subject factors. The PPD was analyzed in the same way (see Table 1).

Since eight ANOVA sets were conducted in the current study, the Benjamini–Hochberg method was used to correct for false discovery rate (FDR) (Benjamini & Hochberg 1995). The ANOVAs yielded a total of 56 *p* values, which were included in

TABLE 1. Overview over the conducted ANOVA sets and all significant main and interaction effects

ANOVA	Within-subject Factors	Effect	*p*	η^2
Identification performance	3 × List length 2 × Task difficulty predictability 2 × Noise reduction	List length Noise reduction	0.003 <0.001	0.22 0.70
Recall performance	3 × List length 2 × Task difficulty predictability 2 × Noise reduction	List length Noise reduction	<0.001 <0.001	0.92 0.59
Sentence baseline 3 sentences/lists	2 × Task difficulty predictability 2 × Noise reduction 3 × Serial position	Task difficulty predictability Serial position	0.019 0.002	0.19 0.25
Sentence baseline 5 sentences/lists	2 × Task difficulty predictability 2 × Noise reduction 5 × Serial position	Task difficulty predictability Serial position Task difficulty predictability × serial position	0.019 <0.001 0.007	0.22 0.30 0.14
Sentence baseline 7 sentences/lists	2 × Task difficulty predictability 2 × Noise reduction 7 × Serial position	Task difficulty predictability Serial position Noise reduction Task difficulty predictability × serial position Noise reduction × serial position	<0.001 <0.001 0.031 <0.001 0.005	0.43 0.40 0.19 0.28 0.13
PPD 3 sentences/lists	2 × Task difficulty predictability 2 × Noise reduction 3 × Serial position	Task difficulty predictability Serial position	0.019 <0.001	0.22 0.35
PPD 5 sentences/lists	2 × Task difficulty predictability 2 × Noise reduction 5 × Serial position	Serial position Task difficulty predictability × serial position	<0.001 0.018	0.29 0.12
PPD 7 sentences/lists	2 × Task difficulty predictability 2 × Noise reduction 7 × Serial position	Serial position Task difficulty predictability × serial position	<0.001 <0.001	0.21 0.21

The p values that survived the FDR correction are underlined.
FDR, false discovery rate; PPD, peak pupil dilation.

the correction. All p values below 0.05 that did not survive the FDR correction will be reported. Post hoc pairwise comparisons of interaction effects were corrected for multiple comparisons at the 0.05 level using the Bonferroni method.

RESULTS

Table 1 provides an overview of the conducted ANOVA sets and all significant main effects and interaction effects. Only the p values that are underlined in the table survived the FDR correction.

Identification Performance

The ANOVA investigating identification performance resulted in a significant main effect of list length, $F_{(2,46)} = 4.98$, $p = 0.004$, $\eta^2 = 0.18$. Post hoc pairwise comparisons showed that significantly more words were repeated when lists contained three (93.5%) sentences compared to lists containing seven sentences (91.7%, $p = 0.018$), but not five sentences (91.8%, $p = 0.80$). Furthermore, a significant main effect of noise reduction was found, $F_{(1,23)} = 79.61$, $p < 0.001$, $\eta^2 = 0.78$, indicating that significantly more words were repeated when noise reduction was on (96.0%), compared to when noise reduction was off (88.7%).

Recall Performance

The ANOVA investigating recall performance revealed a significant main effect of list length on recall performance, $F_{(2,46)} = 271.01$, $p < 0.001$, $\eta^2 = 0.92$, indicating that the proportion of recalled words decreased significantly with increasing list length (three sentences/list = 93.0%, five sentences/list = 71.6%, seven sentences/list = 58.2%). Furthermore, a significant main effect of noise reduction on recall performance was found, $F_{(1,23)} = 33.57$, $p < 0.001$, $\eta^2 = 0.59$. Recall performance was better with noise reduction on (76.6%) compared to when noise reduction was off (71.9%).

Sentence baseline dilation

The three sets of ANOVAs conducted on sentence baseline dilation showed that there was a significant main effect of task difficulty predictability for lists of three, $F_{(1,23)} = 5.44$, $p = 0.029$, $\eta^2 = 0.19$, five, $F_{(1,23)} = 6.41$, $p = 0.019$, $\eta^2 = 0.22$, and seven sentences, $F_{(1,23)} = 17.26$, $p < 0.001$, $\eta^2 = 0.43$, indicating that sentence baseline dilation was significantly higher when task difficulty was unpredictable compared to predictable. However, for lists of three and five sentences, this main effect did not survive the FDR correction.

Furthermore, a significant main effect of serial position was found for lists of three, $F_{(2,46)} = 7.47$, $p = 0.002$, $\eta^2 = 0.25$, five, $F_{(4,92)} = 9.89$, $p < 0.001$, $\eta^2 = 0.30$, and seven sentences, $F_{(6,138)} = 15.0$, $p < 0.001$, $\eta^2 = 0.40$. The overall pattern for all list lengths revealed that sentence baseline dilation increased from the first to the second serial position and then decreased again at the third serial position. After the third serial position, the sentence baseline dilation gradually increased over the course of the remaining serial positions in lists of five and seven sentences.

The ANOVA also resulted in a significant two-way interaction between task difficulty predictability and serial position for lists of five, $F_{(4,92)} = 3.75$, $p = 0.007$, $\eta^2 = 0.14$, and seven sentences, $F_{(6,138)} = 9.01$, $p < 0.001$, $\eta^2 = 0.28$. Post hoc pairwise comparisons revealed that the sentence baseline dilation was significantly higher when task difficulty was unpredictable compared to predictable for all serial positions except the first one. Figure 3 depicts the baseline pupil dilation at each serial position for each list length when task difficulty is predictable and unpredictable. The significant differences obtained from the post hoc analysis are marked on the figure.

A significant main effect of noise reduction was found for lists of seven sentences, $F_{(1,23)} = 5.26$, $p = 0.031$, $\eta^2 = 0.19$, indicating that sentence baseline dilation was higher with noise reduction on compared to off. This main effect did not survive the FDR correction. However, a significant two-way interaction between serial position and noise reduction was found, $F_{(6,138)} = 3.28$, $p = 0.005$, $\eta^2 = 0.13$, which indicated that sentence baseline

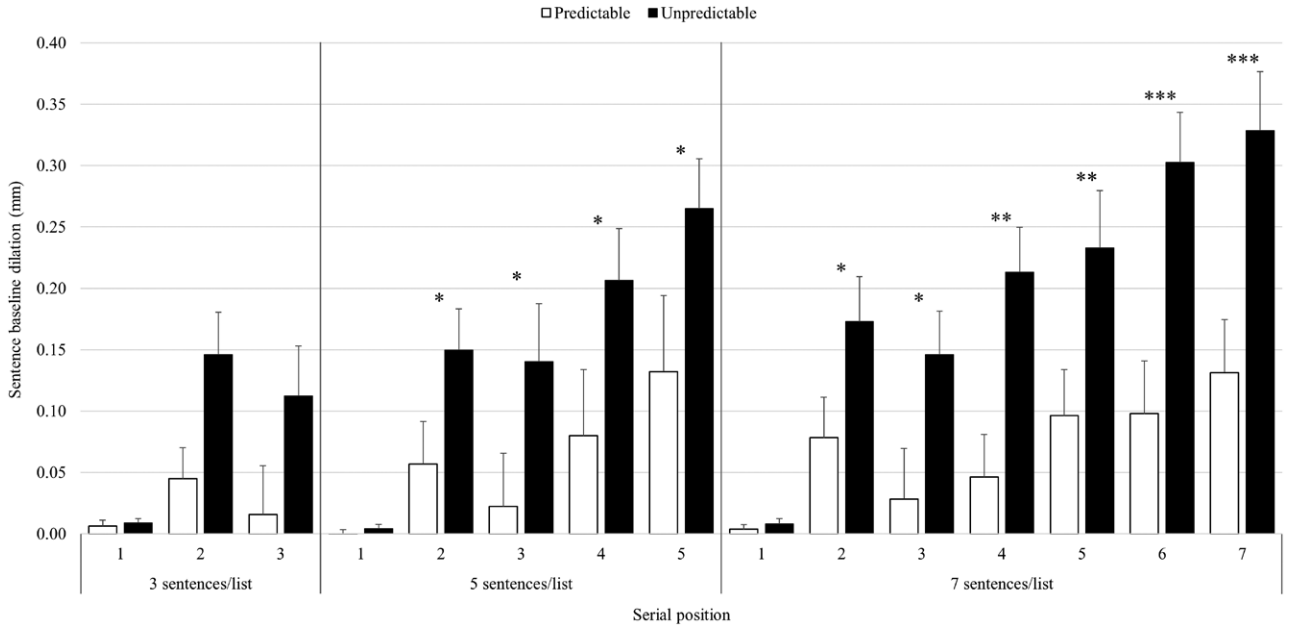

Fig. 3. Sentence baseline dilation per serial position for predictable and unpredictable task difficulty for all list lengths. The significant two-way interaction effects between task difficulty predictability and serial position are indicated (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$). The error bars show the standard error.

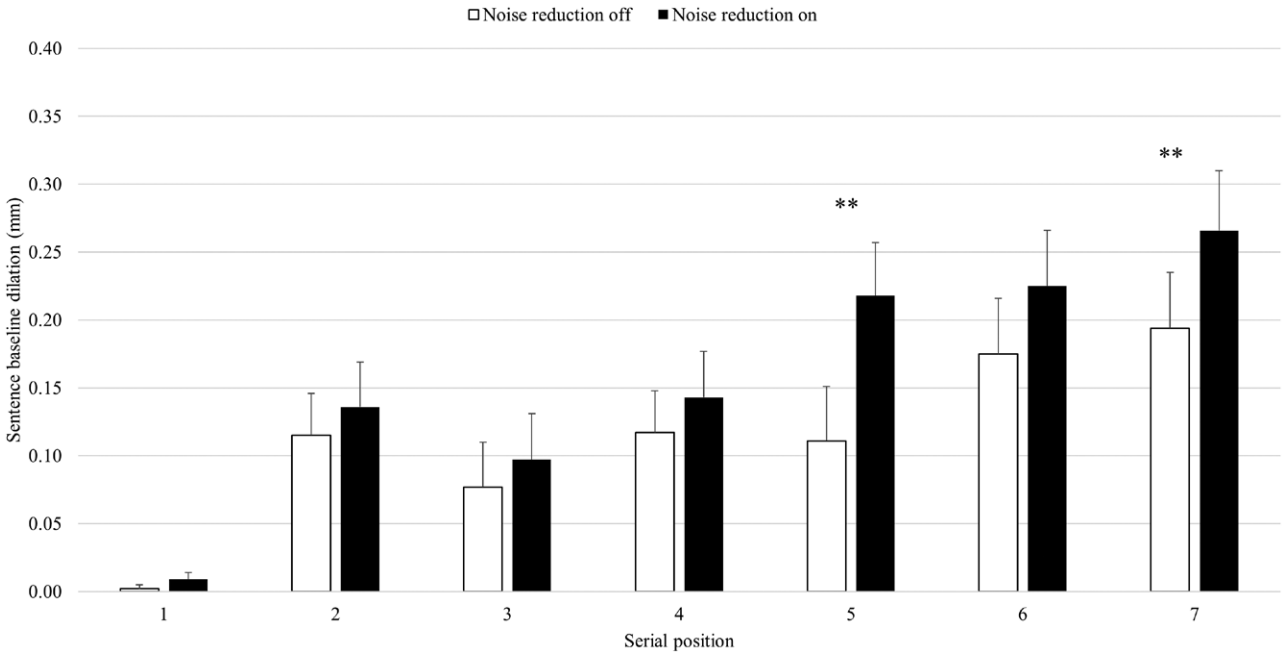

Fig. 4. Two-way interaction effect between noise reduction and serial position on sentence baseline dilation for lists of seven sentences (**$p < 0.01$). The error bars show the standard error.

dilation was significantly higher with noise reduction on compared to off for the fifth and seventh serial position (Fig. 4).

Peak Pupil Dilation

A significant main effect of task difficulty predictability on PPD was found for lists of three sentences, $F_{(1,23)} = 6.38$, $p = 0.019$, $\eta^2 = 0.22$, indicating that PPD was larger when task difficulty was unpredictable compared to predictable. However, this main effect did not survive the FDR correction.

Furthermore, the three sets of ANOVAs on PPD revealed a significant main effect of serial position for lists of three, $F_{(2,46)} = 12.10$, $p < 0.001$, $\eta^2 = 0.35$, five, $F_{(4,92)} = 9.40$, $p < 0.001$, $\eta^2 = 0.29$, and seven, $F_{(6,138)} = 12.46$, $p < 0.001$, $\eta^2 = 0.21$, sentences per list. The overall pattern showed that PPD was highest at the first serial position, after which it decreased and remained relatively stable over the course of the remaining serial positions.

A significant two-way interaction between task difficulty predictability and serial position was found for lists of five,

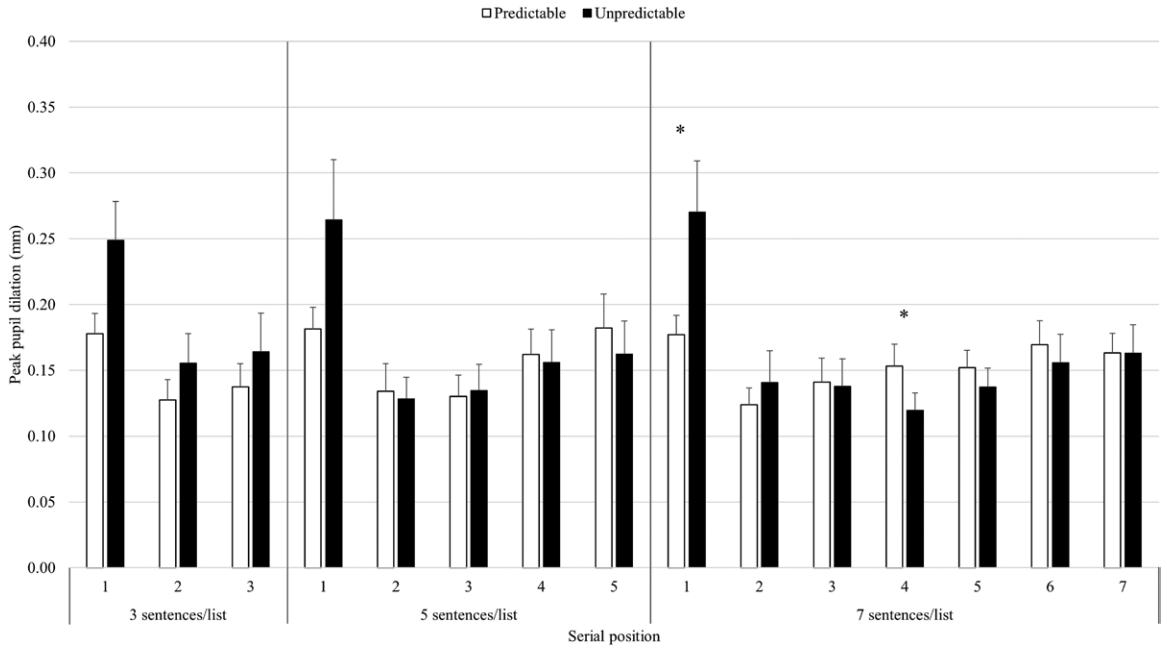

Fig. 5. Peak pupil dilation per serial position for predictable and unpredictable task difficulty for all list lengths. The significant two-way interaction effects between task difficulty predictability and serial position are indicated (*$p < 0.05$). The error bars show the standard error.

$F_{(4,92)} = 3.16$, $p = 0.018$, $\eta^2 = 0.12$, and seven, $F_{(6,138)} = 6.00$, $p < 0.001$, $\eta^2 = 0.21$, sentences. In the case of lists of five sentences, this interaction did not survive the FDR correction. For lists of seven sentences, post hoc pairwise comparisons showed that PPD was significantly larger when task difficulty was unpredictable compared to predictable at the first serial position. At the fourth serial position, a small yet significant difference was found, such the PPD was larger when task difficulty was predictable compared to unpredictable. Figure 5 depicts the PPD at each serial position for each list length when task difficulty is predictable and unpredictable. The significant differences indicated by the post hoc analyses are marked on the figure.

DISCUSSION

The aims of the current study were to investigate the effects of task difficulty predictability and noise reduction on WM resource allocation using the SWIR test and pupillometry. Moreover, we wanted to investigate whether the behavioral outcomes of the SWIR test and physiological responses of pupillometry reflect the same cognitive processes related to WM resource allocation.

Effects of Task Difficulty Predictability on Resource Allocation

As expected, task difficulty predictability did not have an effect on recall performance in the SWIR test, so that the proportion of recalled words was similar independently of whether task difficulty was predictable or unpredictable. This corroborates findings by Micula et al. (2020) and suggests that task difficulty predictability does not influence WM resource allocation to either speech processing or storage. This finding is also in line with the hypothesis related to the first aim of the current study.

Interestingly, task difficulty predictability did have an effect on pupil dilation responses, especially on sentence baseline dilation. This finding revealed that, although there was no significant difference at the first serial position, sentence baseline dilation was significantly higher throughout lists of five and seven sentences when task difficulty was unpredictable compared to predictable. Based on our hypothesis, it is unlikely that the higher sentence baseline dilation when task difficulty was unpredictable reflects an increase in resource allocation to processing or storage of speech, as neither identification nor recall performance were better in this condition compared to when task difficulty was predictable. As mentioned previously, Micula et al. (2020) found that, although task difficulty predictability did not affect overall recall performance, participants recalled words in a different order when task difficulty was unpredictable and noise reduction was off. When list length is unknown, it is not possible for the participants to adapt their recall strategy based on expectations regarding task difficulty (Grenfell-Essam & Ward 2012). Thus, the sentence baseline dilation may reflect differences in recall strategy employed when task difficulty is predictable or unpredictable. Furthermore, baseline pupil dilation is also believed to reflect responses to arousal, attention and engagement in addition to other cognitive processes (Beatty 1982; Laeng et al. 2012). In a recent study, increased baseline dilation in individuals with poor hearing has been interpreted as increased attentional allocation or arousal reflecting task anxiety, that is, insecurity regarding likely success (Ayasse & Wingfield 2020). Unsworth and Robison (2014), who used a change detection task with colored squares in their study, provided evidence that baseline dilation may be modulated by attention allocation. They argue that baseline pupil diameter is smaller before trials during which participants are less attentive. Thus, the baseline dilation is considered to provide an index of change in alertness or engagement over the course of the task. Due to the uncertainty caused by the unpredictable task difficulty in the present study, the increasing sentence baseline dilation may reflect arousal associated with task anxiety in anticipation of the end of the list or an increasing level of alertness or task engagement throughout the list (Unsworth & Robison 2014; Ayasse & Wingfield 2020). The effect of task difficulty predictability on sentence baseline dilation is in line with the expectations presented for the first aim of the current study, although no effect was found for lists of three sentences. These lists are presumably too short to result in an accumulation of arousal or alertness/task engagement.

The effect of task difficulty predictability on PPD was more limited than on sentence baseline dilation, mainly indicating that the PPD was significantly larger at the first serial position when task difficulty was unpredictable compared to predictable. Regarding the first serial position, Zhang et al. (2021) obtained a similar pattern in their study, the PPD being larger compared to the rest of the positions. This was interpreted as a larger amount of resources being allocated to speech recognition at the beginning of the list. Since no studies have investigated task difficulty predictability as was done in the current study, it is not clear why this pattern at the first serial position was more prominent when task difficulty was unpredictable. Although this interaction only survived the FDR correction for lists of seven sentences, the same pattern was observed for the shorter lists as well. We speculate that the difference between the two conditions at the beginning of the list may arise due to differences in instruction. When task difficulty was predictable, the participants were told before each list how many sentences they would hear, thus preparing them for the start of the list, while the participants did not get any verbal cues before the list when task difficulty was unpredictable. In the latter condition, the beginning of the list was only signaled by the onset of the competing speech, which may have triggered an initial increased allocation of resources to processing of speech, after which it decreased and stabilized.

Effects of Noise Reduction on Resource Allocation

Studies using the SWIR test (Ng et al. 2013, 2015; Micula et al. 2020) have provided evidence showing that noise reduction attenuates the competing speech, leading to easier segregation of the target speech and thus facilitating the RAMBPHO function and implicit processing, as described in the ELU model (Rönnberg et al. 2008, 2013, 2019). Consequently, when noise reduction is on, less processing resources are needed to understand heard speech, increasing the amount of resources available for storage, which is reflected in better recall performance in the SWIR test. Thus, the significant improvement in recall performance with noise reduction on is in line with the hypothesis presented in the second aim of the current study and further supports evidence that noise reduction can lower the need for explicit processing in noisy environments.

Although Bönitz et al. (2021) found that task-evoked mean pupil dilation was lower when noise reduction was on compared to off, they found no significant effect of noise reduction

on sentence baseline dilation. Our findings indicate that noise reduction has an effect on sentence baseline dilation but not on the task-evoked PPD. This difference may arise due to the differences in hearing status of the participant groups, the inclusion of longer lists in the current study or the difference in task-evoked pupillary response calculations. Our findings showed that sentence baseline dilation was overall higher with noise reduction on compared to off for all serial positions, the difference being significant toward the end of lists of seven sentences. This is consistent with the improvement in recall performance with noise reduction on. Recent studies indicate that baseline dilation reflects the amount of resources allocated for maintenance of speech in memory for subsequent recall (Zhang et al. 2021; Bönitz et al. 2021). The findings of the current study seem to corroborate this idea. Furthermore, the increase in sentence baseline dilation with noise reduction on compared to off towards the end of the list demonstrates that noise reduction increases the capacity to maintain more words in memory. Based on prior findings suggesting that noise reduction facilitates implicit processing and thereby frees up resources to be used for storage (Ng et al., 2013, 2015; Micula et al. 2020), higher sentence baseline dilation alongside better recall performance may reflect a higher degree of implicit processing (Rönnberg et al. 2008, 2013, 2019). Alternatively, the higher sentence baseline dilation may reflect the strategies that are applied to maintain the level of attention allocation or task engagement (Unsworth & Robison 2014) as the task demands increase (i.e., keeping more words in memory over the course of the list), which may in turn increase the resources devoted to maintaining more words in memory. We predicted a larger sentence baseline dilation with noise reduction on in the second aim of the current study. However, this was only found for lists of seven sentences, which may be due to the fact that this list length poses the highest task demands in terms of memory.

SWIR Test and Pupillometry as Complementary Measures

To summarize, the findings of the current study demonstrate that although task difficulty predictability has no effect on recall performance, it does have an effect on pupil dilation responses. The sentence baseline dilation for longer lists of five and seven sentences was significantly higher when task difficulty was unpredictable compared to predictable, except at the first serial position (Fig. 3). For lists of seven sentences, the PPD at the first serial position was significantly higher when task difficulty was unpredictable compared to predictable, indicating an increase in allocation of resources to speech processing at the beginning of the list when the list length cue was absent (Fig. 5). Recall performance was overall significantly better and sentence baseline dilation for lists of seven sentences was significantly higher, especially towards the end of the list, when noise reduction was on compared to off (Fig. 4). These findings indicate that, while some cognitive processes may be reflected in both behavioral performance and pupillary responses, this is not always the case.

Recent studies have provided evidence that the tonic baseline dilation reflects the memory load over the course of a task, that is, the amount of WM resources allocated to rehearsing or maintaining words in memory for subsequent recall (Zhang et al. 2021; Bönitz et al. 2021). When it comes to the effect of noise reduction on resource allocation, that is, the increased capacity to allocate resources for storage when noise reduction is on, it seems to be reflected both in the behavioral performance and pupillary responses. Although the two measures seem to capture the same processes related to WM resource allocation, pupillary responses are a valuable complement to the behavioral performance due to the temporal information it can provide. As was discussed previously, pupillometry is an online measure, while the SWIR test is an offline measure of resource allocation. Consequently, recall performance in the SWIR test reveals that more resources are allocated for storage of speech when noise reduction is on. However, tracking the sentence baseline dilation across serial positions reveals that the allocation of resources for storage increases as the memory load increases. Furthermore, a higher amount of resources is allocated for storage when noise reduction is on compared to off as memory load increases over the course of the list.

On the other hand, while the effect of task difficulty predictability seems to be captured by pupillometry, it is not reflected in overall recall performance in the SWIR test. This suggests that the effect of noise reduction and the effect of task difficulty predictability are driven by different mechanisms. We speculate that the different outcomes of the SWIR test and pupillometry in terms of task difficulty predictability may be due to the fact that the SWIR test measures WM resource allocation to storage of speech under varying processing demands (Ng et al. 2013, 2015; Lunner et al. 2016; Micula et al. 2020), while pupil dilation responses may also reflect other cognitive processes (Aston-Jones & Cohen 2005; Laeng et al. 2012; Peelle 2018). Thus, while noise reduction seems to affect cognitive processes reflected in the recall performance and pupil dilation responses, task difficulty predictability seems to mainly affect pupil dilation responses, presumably reflecting differences in recall strategy (Micula et al. 2020), online changes in task engagement/alertness over the course of the task or arousal associated with task anxiety (Unsworth & Robison 2014; Ayasse & Wingfield 2020).

The current study demonstrates that pupillary responses may provide information about different systems or mechanisms that drive certain responses. For instance, the sentence baseline dilation was interpreted as an index of recall strategy, task engagement/alertness or arousal when task difficulty predictability was manipulated, while the same response was interpreted as an index of resource allocation to storage or degree of implicit processing when noise reduction was manipulated. To overcome this challenge, it is necessary to design well-controlled experimental contrasts in order to facilitate interpretation. Behavioral responses, such as recall performance in this case, can be essential for untangling underlying mechanisms which result in similar patterns of pupillary responses. Hence, the offline behavioral performance obtained in the SWIR test is a suitable measure to supplement online physiological pupillometry data.

Limitations

In the current study we administered the Montreal Cognitive Assessment (MoCA), which is a short cognitive screening tool (Nasreddine et al. 2005). The mean MoCA score was 26.6 points (SD = 2.00). Although Nasreddine et al. (2005) recommend a cutoff score of 26 out of 30 points, some studies indicate that this criterion may be too strict, leading to an inflated false positive rate for older adults or individuals with lower education (Carson et al. 2018). Carson et al. recommend a cutoff score of

23 out of 30 points to compensate for this issue. One participant scored below the cutoff score of 23, which is a potential limitation. However, the respective participant did not report any difficulties completing the SWIR test and their performance was similar to other participants.

CONCLUSION

The findings of the current study demonstrate that task difficulty predictability did not have an effect on recall performance in the SWIR test. However, although there was no difference in sentence baseline dilation at the first serial position, sentence baseline dilation was higher over the course of the list when task difficulty was unpredictable compared to predictable. The lack of effect of task difficulty predictability on both identification and recall performance suggests that task difficulty predictability does not affect WM resource allocation to processing or storage of speech. We argue that when task difficulty is unpredictable, the sentence baseline dilation reflects a different recall strategy (Micula et al. 2020) or an accumulation of alertness/task engagement or arousal reflecting task anxiety (Unsworth & Robison 2014; Ayasse & Wingfield 2020).

Furthermore, noise reduction led to better recall performance when it was on. This finding demonstrates that noise reduction can reduce the amount of WM resources needed to process speech, thus freeing up resources to be used for storage (Ng et al. 2013, 2015; Micula et al. 2020). This was also reflected in a higher sentence baseline dilation when noise reduction was on, especially toward the end of longer lists when more words needed to be maintained in memory. We argue that the higher sentence baseline dilation in conjunction with better recall performance with noise reduction on is an index of the increased capacity to allocate WM resources to storage of speech (Zhang et al. 2021; Bönitz et al. 2021). No effect of noise reduction on PPD was found, suggesting that adding a memory task to a speech understanding task overwrites the well-known effect of perceptual factors on PPD (Zekveld et al. 2018b; Zhang et al. 2021).

The effect of task difficulty predictability was only captured by pupillary responses, while the effect of noise reduction was captured by both behavioral performance and pupillary responses. The SWIR test is a measure of WM resource allocation, while pupillometry may reflect additional cognitive processes, which may explain why the effect of task difficulty predictability was only captured by the latter measure. However, since the SWIR test is an offline measure and pupillometry is an online measure, they reflect different temporal aspects of WM resource allocation. Furthermore, the findings of the current study demonstrate that similar pupil dilation patterns may be triggered by different mechanisms, that is, the effects of task difficulty predictability and noise reduction on sentence baseline dilation. This highlights that behavioral performance can be essential for interpretation of pupil dilation responses. In sum, the SWIR test and pupillometry provide complementary information on WM resource allocation to processing and storage of speech.

ACKNOWLEDGMENTS

The authors thank Jens-Christian Britze Kijne and Morten Ammekilde Nielsen from Oticon A/S, Smørum, Denmark, for the support with the test setup of this study and participant recruitment, as well as Lu Xia from Oticon A/S, Smørum, Denmark and Pierre-Yves Hasan from Oticon Medical, Smørum, Denmark, for their help with pupil data analysis.

This study is part of a collaborative PhD project between Oticon A/S and Linnaeus Centre HEAD, Linköping University and is funded by the William Demant Foundation.

A.M. designed and conducted the study, performed statistical analysis and wrote the article. L.F. provided support with pupil data extraction. E.H.N.N., L.F., D.W., and J.R. designed the study and provided critical revision. M.C.J. and D.K.L. conducted the study and performed preliminary statistical analysis, which was used for their master thesis at the University of Copenhagen.

The authors have no conflicts of interest to declare.

Address for correspondence: Andreea Micula, Department of Behavioural Sciences and Learning, Linköping University, SE-581 83 Linköping, Sweden. E-mail: andmi366@student.liu.se

Received September 23, 2020; accepted February 21, 2021; published online ahead of print April 15, 2021.

REFERENCES

Arehart, K., Souza, P., Kates, J., Lunner, T., Pedersen, M. S. (2015). Relationship among signal fidelity, hearing loss, and working memory for digital noise suppression. *Ear Hear*, *36*, 505–516.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu Rev Neurosci*, *28*, 403–450.

Ayasse, N. D., & Wingfield, A. (2020). Anticipatory baseline pupil diameter is sensitive to differences in hearing thresholds. *Front Psychol*, *10*, 1–7.

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Ann Rev Psychol*, *63*, 1–29.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol Bull*, *91*, 276–292.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J R Stat Soc*, *57*, 289–300.

Bönitz, H., Lunner, T., Finke, M., Fiedler, L., Lyxell, B., Riis, S. K., Ng, E., Valdes, A. L., Büchner, A., Wendt, D. (2021). How do we allocate our resources when listening and memorizing speech in noise? A pupillometry study. *Ear Hear, 42,* 846–859.

Buus, S., & Florentine, M. (2001). Growth of loudness in listeners with cochlear hearing losses: recruitment reconsidered. *J Assoc Res Otolaryngol*, *03*, 120–139.

Carson, N., Leach, L., Murphy, K. J. (2018). A re-examination of montreal cognitive assessment (MoCA) cutoff scores. *Int J Geriatr Psychiatry*, *33*, 379–388.

Dryden, A., Allen, H. A., Henshaw, H., Heinrich, A. (2017). The association between cognitive performance and speech-in-noise perception for adult listeners: a systematic literature review and meta-analysis. *Trends Hear*, *21*, 2331216517744675.

Edwards, B. (2016). A model of auditory-cognitive processing and relevance to clinical applicability. *Ear Hear*, *37* (Suppl 1), 85S–91S.

Gordon-Salant, S., & Cole, S. S. (2016). Effects of age and working memory capacity on speech recognition performance in noise among listeners with normal hearing. *Ear Hear*, *37*, 593–602.

Gosselin, P. A., & Gagné, J. P. (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *Int J Audiol*, *50*, 786–792.

Grenfell-Essam, R., & Ward, G. (2012). Examining the relationship between free recall and immediate serial recall: The role of list length, strategy use, and test expectancy. *J Mem Lang*, *67*, 106–148.

Jensen, J., & Pedersen, M. S. (2015). Analysis of beamformer directed single-channel noise reduction system for hearing aid applications. *Proceedings of the EEE International Conference on Acoustics, Speech and Signal Processing*, *Australia*, 5728–5732.

Kahneman, D. (1973). *Attention and Effort*. Prentice-Hall.

Kjems, U., & Jensen, J. (2012). Maximum likelihood based noise covari- ance matrix estimation for multi-microphone speech enhancement. *Proceedings of the European Signal Processing Conference*, *Romania*. 295–299.

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

Laeng, B., Sirois, S., Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspect Psychol Sci*, *7*, 18–27.

Lemke, U., & Besser, J. (2016). Cognitive load and listening effort: concepts and age-related considerations. *Ear Hear*, *37* (Suppl 1), 77S–84S.

Lunner, T., Rudner, M., Rönnberg, J. (2009). Cognition and hearing aids. *Scand J Psychol*, *50*, 395–403.

Lunner, T., Rudner, M., Rosenbom, T., Ågren, J., Ng, E. H. (2016). Using speech recall in hearing aid fitting and outcome evaluation under ecological test conditions. *Ear Hear*, *37* (Suppl 1), 145S–154S.

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *Q J Exp Psychol A*, *58*, 22–33.

Micula, A., Ng, E. H. N., El-Azm, F., Rönnberg, J. (2020). The effects of noise reduction, background noise and task difficulty on recall. *Int J Audiol*, *59*, 792–800.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., Chertkow, H. (2005). The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *J Am Geriatr Soc*, *53*, 695–699.

Neher, T., Wagener, K. C., Fischer, R. L. (2018). Hearing aid noise suppression and working memory function. *Int J Audiol*, *57*, 335–344.

Ng, E. H., Rudner, M., Lunner, T., Pedersen, M. S., Rönnberg, J. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *Int J Audiol*, *52*, 433–441.

Ng, E. H., Rudner, M., Lunner, T., Rönnberg, J. (2015). Noise reduction improves memory for target language speech in competing native but not foreign language speech. *Ear Hear*, *36*, 82–91.

Nielsen, J. B., & Dau, T. (2011). The Danish hearing in noise test. *Int J Audiol*, *50*, 202–208.

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

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

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear Hear*, *37*, 5S–27S.

Rönnberg, J., Holmer, E., Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *Int J Audiol*, *58*, 247–261.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, O., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., Rudner, M. (2013). The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Front Syst Neurosci*, *7*, 31.

Rönnberg, J., Rudner, M., Foo, C., Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *Int J Audiol*, *47* (Suppl 2), S99–105.

Smith, S. L., Pichora-Fuller, M. K., Alexander, G. (2016). Development of the word auditory recognition and recall measure: A working memory test for use in rehabilitative audiology. *Ear Hear*, *37*, e360–e376.

Unsworth, N., & Robison, M. K. (2014). Individual differences in the allocation of attention to items in working memory: Evidence from pupillometry. *Psychon Bull Rev*, *22*, 757–765.

Wendt, D., Hietkamp, R. K., Lunner, T. (2017). Impact of noise and noise reduction on processing effort: A pupillometry study. *Ear Hear*, *38*, 690–700.

Wingfield, A., Amichetti, N. M., Lash, A. (2015). Cognitive aging and hearing acuity: Modeling spoken language comprehension. *Front Psychol*, *6*, 684.

Winn, M. B., Wendt, D., Koelewijn, T., Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends Hear*, *22*, 2331216518800869.

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

Zekveld, A. A., Kramer, S. E., Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear Hear*, *32*, 498–510.

Zekveld, A. A., Kramer, S. E., Rönnberg, J., Rudner, M. (2018b). In a concurrent memory and auditory perception task, the pupil dilation response is more sensitive to memory load than to auditory stimulus characteristics. *Ear Hear*, *40*, 272–286.

Zhang, Y., Lehmann, A., Deroche, M. (2021). Disentangling listening effort and memory load beyond behavioural evidence: Pupillary response to listening effort during a concurrent memory task. *PLoS ONE*, *16*, e0233251.