cortex 142 (2021) 296–316

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Special Issue "Mapping sound to meaning under challenging conditions": Research Report

The costs (and benefits) of effortful listening on context processing: A simultaneous electrophysiology, pupillometry, and behavioral study

Jack W Silcox [a,*] and *Brennan R. Payne* [a,b]

[a] *Department of Psychology, University of Utah, USA*
[b] *Interdepartmental Neuroscience Program, University of Utah, USA*

A R T I C L E I N F O

Article history:
Received 6 November 2020
Reviewed 5 January 2021
Revised 2 April 2021
Accepted 10 June 2021
Published online 2 July 2021

Keywords:
Listening effort
N400
Pupillometry
Linguistic context
Memory
Single-trial analysis

A B S T R A C T

There is an apparent disparity between the fields of cognitive audiology and cognitive electrophysiology as to how linguistic context is used when listening to perceptually challenging speech. To gain a clearer picture of how listening effort impacts context use, we conducted a pre-registered study to simultaneously examine electrophysiological, pupillometric, and behavioral responses when listening to sentences varying in contextual constraint and acoustic challenge in the same sample. Participants ($N = 44$) listened to sentences that were highly constraining and completed with expected or unexpected sentence-final words ("The prisoners were planning their *escape/party*") or were low-constraint sentences with unexpected sentence-final words ("All day she thought about the *party*"). Sentences were presented either in quiet or with +3 dB SNR background noise. Pupillometry and EEG were simultaneously recorded and subsequent sentence recognition and word recall were measured. While the N400 expectancy effect was diminished by noise, suggesting impaired real-time context use, we simultaneously observed a beneficial effect of constraint on subsequent recognition memory for degraded speech. Importantly, analyses of trial-to-trial coupling between pupil dilation and N400 amplitude showed that when participants' showed increased listening effort (i.e., greater pupil dilation), there was a subsequent recovery of the N400 effect, but at the same time, higher effort was related to poorer subsequent sentence recognition and word recall. Collectively, these findings suggest divergent effects of acoustic challenge and listening effort on context use: while noise impairs the rapid use of context to facilitate lexical semantic processing in general, this negative effect is attenuated when listeners show increased effort in response to noise. However, this effort-induced reliance on context for online word processing comes at the cost of poorer subsequent memory.

* *Corresponding author*. 380 South 1530 East, Salt Lake City, UT, 84106, USA.
E-mail address: jack.silcox@utah.edu (J.W. Silcox).

https://doi.org/10.1016/j.cortex.2021.06.007

1. Introduction

Although listening to speech appears to be simple, there are many factors that can make speech comprehension an inherently difficult process. Perceptual challenge accompanying the speech signal in the form of background noise or hearing impairment can increase the difficulties associated with speech comprehension by increasing the draw on limited cognitive and neural resources available to a listener (Peelle, 2018; Pichora-Fuller et al., 2016). The deliberate allocation of these limited resources to overcome these challenges is referred to as *listening effort* (Pichora-Fuller et al., 2016). Although research on listening effort dates back to the 1960s (e.g., Rabbitt, 1968), the factors underlying effortful listening and its impacts on higher level language comprehension are not well understood. For example, although it has been proposed that listeners can use semantic and syntactic information available in the ongoing linguistic context to mitigate the effects of effortful listening (e.g., Benichov, Cox, Tun, & Wingfield, 2012; Lash, Rogers, Zoller, & Wingfield, 2013; Pichora-Fuller, 2008; Sheldon, Pichora-Fuller, & Schneider, 2008), other work in a growing body of research in the field of cognitive electrophysiology that has shown that a listener's ability to use context to facilitate online word processing (e.g., as reflected by the N400 component of the event-related brain potential, ERP), is reduced when listening is more effortful (e.g., Romero-Rivas, Martin, & Costa, 2016; Schiller et al., 2020). Therefore, the goal of the current study is to use methodologies from both literatures in the same sample to help resolve this apparent discrepancy and gain a clearer picture of how listeners use contextual information while experiencing changes in listening effort.

1.1. *Listening effort*

Recently, a large group of experts in the hearing sciences proposed the Framework for Understanding Effortful Listening (FUEL; Pichora-Fuller et al., 2016) to synthesize work on the cognitive and neural constraints on listening. There are two main components of FUEL: first, the listener's limited pool of neurocognitive resources and, second, the listener's resource allocation policy. The available cognitive resource capacity is modulated by the arousal level of the listener and by the demands placed on the system. The resource allocation policy is guided by automatic and intentional attention and is modulated by general arousal level. Importantly, the FUEL emphasizes that arousal, attention and motivation levels have a strong influence over engagement and the allocation of cognitive resources (Pichora-Fuller et al., 2016).

For example, if a listener is fatigued or finds displeasure in a listening task, they may not allocate sufficient resources for successful processing *regardless of the demands of the task*. Therefore, Pichora-Fuller et al. (2016) define listening effort as "the *deliberate* allocation of mental resources to overcome obstacles in goal pursuit when carrying out … listening tasks" (p. 10S; emphasis added). Important to the current discussion, FUEL predicts that, over the time course of a listening activity, the amount of effort experienced by a listener can vary depending on the demand placed on the system (e.g., how degraded the speech signal is), the arousal level of the listener (e.g., how fatigued they are), and the attention and motivational level of the listener (e.g., how important successful listening is to the listener; see Brehm & Self, 1989; Brehm et al., 1983; Richter, Gendolla, & Wright, 2016).

1.2. *Listening effort and the effects of context on behavioral measures*

The beneficial effects of context on offline measures of word recognition and memory are robust in the speech audiology literature (see Payne & Silcox, 2019 for a recent review). For example, Pichora-Fuller et al. (1995) investigated how context is used in less than ideal listening scenarios by manipulating contextual constraint and level of background noise and asking participants to identify sentence-final words. They found that participants did better at identifying the sentence-final word as the signal-to-noise ratio (SNR) increased (i.e., background noise decreased). However, they also found that there was a benefit from context, in that word recognition performance was better in highly constraining contexts as compared to less constraining contexts. In other words, in the conditions where the speech signal was degraded by competing background noise, participants were better able to identify words that were preceded by a highly constraining context.

Increases in listening effort have also been found to negatively influence memory processes (Guang, Lefkowitz, Dillman-Hasso, Brown, & Strand, 2021; Payne et al., 2021; Piquado, Benichov, Brownell, & Wingfield, 2012; Rabbitt, 1968, 1991). However, there is evidence that these effects are offset by the presence of supportive context (e.g., Gordon-Salant & Fitzgibbons, 1997; McCoy et al., 2005; Winneke, Schulte, Vormann, & Latzel, 2020). For example, Gordon-Salant and Fitzgibbons (1997) presented participants with and without hearing impairments with sentences embedded in 12-talker babble background noise. Half of the sentences were low constraint and half of the sentences were high constraint. Participants were asked to recall what they heard to the best of their ability after each sentence was presented. They found that all participants showed worse free recall when listening to less constraining contexts, particularly for adults with hearing impairment. However, all listeners, regardless of hearing acuity and age, performed at ceiling when listening to sentences with highly constraining context, suggesting that sentential constraint can eliminate the negative effects of hearing loss and noise on subsequent memory.

1.3. *Electrophysiological studies of context use and listening effort*

The beneficial effects of linguistic context on word processing have been observed in the field of cognitive electrophysiology since the 1980's (Kutas & Hillyard, 1980). The N400, the most widely studied language-related ERP component, is a centro-posterior negative deflection that peaks in healthy young adults around 400 msec after the onset of a stimulus and is strongly related to the semantic processing of meaning-bearing stimuli (for detailed reviews of the N400 see Kutas & Federmeier, 2000, 2011). This ERP component is thought to originate from a widely-distributed but left-lateralized semantic network, comprising superior and middle temporal gyrus, angular gyrus, and anterior temporal cortex with additional possible generators in left inferior frontal cortex (for reviews see, Lau, Phillips, & Poeppel, 2008; Van Petten & Luka, 2006). Although the N400 is sensitive to a whole host of factors that impact semantic memory access (see e.g., Kutas & Federmeier, 2011), the amplitude of the N400 is most strongly modulated by the degree with which a word is predicted by the preceding semantic context, i.e., its cloze probability (Kutas & Hillyard, 1984). Therefore, most accounts of the N400 context effect suggest that supportive linguistic contexts facilitate semantic memory-related processes (Kutas & Federmeier, 2011).

Although the N400 has been used extensively to study the way in which contextual information is used in ideal listening scenarios, only a small number of studies have begun to explore how acoustic challenge may influence the use of context. Obleser and Kotz (2011) experimentally manipulated stimulus degradation via noise vocoding (see Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) and found that the N400 mean amplitude decreased and peak latency increased as speech intelligibility decreased (see also Strauß, Kotz, & Obleser, 2013; Aydelott, Dick, & Mills, 2006). Similarly, Romero-Rivas et al. (2016) reported findings in which there was a reduced N400 context effect when listening to foreign-accented speech, which is also theorized to induce listening effort (see also Goslin, Duffy, & Floccia, 2012; Schiller et al., 2020). Collectively, these studies suggest that a listener's ability to use context to facilitate online lexical semantic processing may be compromised when listening to perceptually challenging speech. It should be noted however that none of these prior studies have only manipulated intelligibility and *assumed* that listening effort has increased, and so it is difficult to delineate whether these effects arise primarily due to increases in listening effort or directly from the acoustic challenge associated with noise masking or listening to foreign accented speech.

1.4. *Pupillometry and listening effort*

Pupillometry is the measure of changes in pupil size over time. It has been known for some time that there are changes in pupil size related to cognitive processes under constant lighting conditions (Berrien & Huntington, 1943; Hess & Polt, 1960, 1964; for a review of the history of the use of pupillometry in cognitive research see; Sirois & Brisson, 2014). Cognitive-evoked dilations seen in pupillometry have been largely attributed to activity in the locus coeruleus-norepinephrine (LC-NE) system (Breton-Provencher & Sur, 2019; Joshi, Li, Kalwani, & Gold, 2016; Murphy, O'connell, O'sullivan, Robertson, & Balsters, 2014; Reimer et al., 2016; Varazzani, San-Galli, Gilardeau, & Bouret, 2015) but there has been emerging evidence that other midbrain structures, including the pretectal olivary nucleus and the superior colliculus, may also be involved in the cognitive-evoked pupillary response (for a recent review of the neurophysiology of this response see, Joshi & Gold, 2020).

Under constant lighting conditions, pupillometry has been shown to be sensitive to changes in cognitive effort (Hess & Polt, 1964; Sirois & Brisson, 2014; Van Gerven, Paas, Van Merriënboer, & Schmidt, 2004), motivation (Knapen et al., 2016) and arousal (Blackburn & Schirillo, 2012; Bradley, Miccoli, Escrig, & Lang, 2008; Webb et al., 2009). Importantly, pupillometry has started to be utilized with some regularity to study listening effort in speech comprehension (Koelewijn et al, 2012a, 2015; McGarrigle, Dawes, Stewart, Kuchinsky, & Munro, 2017; Zekveld et al., 2010, 2011). Indeed, the tight link between the pupillary response and the LC-NE system (Aston-Jones & Cohen, 2005; Joshi et al., 2016; Reimer et al., 2016) and the importance of arousal in models of listening effort (Peelle, 2018; Pichora-Fuller et al., 2016), make pupillometry an ideal candidate to be an online physiological measure of listening effort. Across studies, there is a reliable pattern of increasingly larger evoked pupillary responses to speech as it becomes increasingly degraded (Koelewijn et al., 2012b, 2014, 2014; Wagner, Toffanin, & Başkent, 2016; Winn, 2016; Winn, Edwards, & Litovsky, 2015; Zekveld et al., 2011, 2013, 2014a, 2014b). At the same time, there is evidence that the relationship between pupil size and intelligibility is nonlinear (McMahon et al., 2016; Ohlenforst et al., 2017; Wendt, Koelewijn, Książek, Kramer, & Lunner, 2018; Zekveld & Kramer, 2014). For example, Wendt et al. (2018) presented listeners with sentences that continuously varied in intelligibility as a function of performance on an immediate sentence recall task. They found that as intelligibility decreased, there was a subsequent decrease in performance and a concomitant increase in pupil size up to a certain threshold. Once performance decreased below 10% accuracy, the pupillary response also decreased. Peak pupillary responses were found for sentences in SNR conditions with 30–70% accuracy, leading to an inverted-U function between pupil dilation and performance (see also Ohlenforst et al., 2017; Zekveld & Kramer, 2014). Wendt et al. (2018) concluded that this pupillary response followed a pattern that would be predicted by models of listening effort (e.g., FUEL, Pichora-Fuller et al., 2016): as input demands increased, so did the effort required for successfully performing the task, as measured by an increase in the pupillary response. However, as speech became increasingly unintelligible, the likelihood of failure even at high levels of effort increased, leading to decreased motivation, and attention likely diverted resources elsewhere, leading to a reduction in effort (indicated by a decrease in the pupillary response). Indeed, after an extensive review of 146 studies looking at the pupil dilation response to auditory stimuli, Zekveld, Koelewijn, and Kramer (2018) concluded that "the pupil response, and the allocation processes reflected by the response, indexes a complex mechanism underlying cognitive resource allocation" and this response "sensitively reflects differences in arousal" (p. 17).

1.5. *The current study*

As can be seen in the preceding review, there are some conflicting findings between the behavioral evidence seen in the field of cognitive audiology and the electrophysiological evidence seen in the field of cognitive neuroscience (for a more thorough discussion see Payne & Silcox, 2019). Although this empirical evidence is not necessarily irreconcilable, the two fields have been somewhat siloed from each other and have come to very different conclusions about how linguistic context is *generally* used when speech is acoustically challenging. For example, in cognitive audiology, linguistic context is often referred to as "supportive" and listeners can "deploy [sentential] context to *compensate* for listening challenges" (Pichora-Fuller, 2008, p. S75, emphasis added). McCoy and colleagues wrote that semantic contextual constraints "*reduce* the perceptual burden on the listener's processing resources … [leaving] *more* resources available for encoding … words in memory, resulting in more successful recall" (2005, p. 31, emphasis added). Therefore, in audiology, it is often implied or explicitly stated that the ability to use linguistic context is not only intact when listening to perceptually challenging speech but that using context can free up resources and can be successfully relied upon to overcome the effects of listening effort. On the other hand, when looking at electrophysiological evidence, Strauβ and colleagues wrote that "… perceptual load … *limits resources* a listener has available for forming predictions as the sentence unfolds" (2013, p. 1393, emphasis added). When seeing a reduction in the N400 context effect when participants were listening to foreign-accented speech, Romero-Rivas and colleagues wrote that "these observations could be explained by *narrowed* lexical expectations" (2016, p. 254, emphasis added). In the field of cognitive electrophysiology, the N400 evidence in particular (which has been used for decades as a valid and reliable online measure of the use of linguistic context, see Kutas & Federmeier, 2011), has led researchers to conclude that the ability to use linguistic context when listening to challenging speech is limited by the increased perceptual load and a strain on available resources. Evidence from both fields, when independently assessed, has led to different broad conclusions about how linguistic context is used when listening to perceptually challenging speech. Moreover, the majority of this past work has assumed increased listening effort under acoustically challenging conditions but did not independently assess listening effort, for example, via pupillometry. This is important because, while acoustic challenge increases cognitive demand, it is not the only factor determining effortful listening. Rather, listening effort reflects the deliberate allocation of cognitive and neural resources in response to increased acoustic challenge, which can vary substantially within a given listening situation (Winn & Teece, 2021; Zekveld et al, 2010, 2018).

Therefore, the goals of this study were to begin to bridge the gap between these fields and better understand the roles of listening effort and context use in challenging listening scenarios. To do this, we utilized methodologies and outcomes used in prior work in cognitive audiology and cognitive electrophysiology in the same sample. Specifically, we simultaneously examined behavioral (e.g., memory) and neural (e.g., ERP) responses to acoustic challenge in speech processing while participants listened to sentences that varied in contextual constraint and lexical expectancy (e.g., Federmeier, Wlotko, De Ochoa-Dewald, & Kutas, 2007; Ng, Payne, Stine-Morrow, & Federmeier, 2018). In addition, we simultaneously recorded pupillometry as an objective and online physiological measure of listening effort. Critically, to directly relate noise-induced listening effort to comprehension and memory processes, we examined the trial-to-trial relationships between variability in task-evoked pupil dilation (as a marker of trial-to-trial variation in listening effort) and both electrophysiological responses and memory measures. By using an online measure of listening effort, we aimed to be able to better understand not just how listening in acoustically challenging scenarios affects the use of context generally, but also how trial by trial dynamic changes in listening effort affect the online and offline use of context.

2. Material & methods

2.1. Preregistration

The current study was preregistered on the Open Science Framework website (https://osf.io/5kmbh). Throughout the remainder of this document, we will be explicit in which hypotheses and analyses were confirmatory (pre-registered) and which were exploratory (Nosek, Ebersole, DeHaven, & Mellor, 2018). All deviations from the pre-registered procedures and analysis plans are transparently reported. Stimuli can be found at: https://osf.io/tv8y6/. Data used in analysis can be found at: https://osf.io/hcrv6/files/. R code used for analyses can be found at: https://osf.io/e7ztg/. We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

2.2. Participants

Informed consent was obtained for forty-four adults[1] (23 female, mean age = 20.6 years, range = 18–34) from the University of Utah community who participated in the experiment for course credit or payment. All were right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield, 1971; see: https://www.brainmapping.org/shared/Edinburgh.php) and had no prior history of neurological issues. All participants had their hearing acuity assessed using pure tone audiometry and speech reception threshold tests in each ear via a modified Hughson–Westlake pure tone identification procedure. No participants had any identifiable hearing impairment. For more details on the assessments used and their outcomes see: https://osf.io/3u65g/. Each participant performed a modified "FAS" phonemic fluency test (Benton et al., 1978) , in which they were asked to name as many words that begin with the letter "F" as quickly as they could in 60 s while not repeating words or using proper nouns. Additionally, they completed a short-form computerized version of the reading span task (Oswald, McAbee, Redick, & Hambrick, 2015; see: https://englelab.gatech.edu/

[1] An a priori power analysis (using PANGEA; Westfall, 2015) suggested that with a sample size of N = 48, we would have a power of .827 to detect a standardized effect size of .25 or less, assuming alpha = .05. However, due to the COVID-19 pandemic, our data collection was stopped 4 participants short of our original goal. With a sample size of 44, our a priori power would be reduced to .793.

complexspantasks), and an extended range vocabulary test (Ekstrom, Dermen, & Harman, 1976; Tombaugh, Kozak, & Rees, 1999; Payne et al., 2015; Oswald et al., 2015). Legal copyright restrictions prevent public archiving of the extended range vocabulary test used in this study, which can be obtained from the copyright holders (see Ekstrom et al., 1976). For more detail on the outcomes of these neuropsychological assessments and other details on demographic information see: https://osf.io/3u65g/(note that this document does not contain the materials referenced, only their outcomes). All participants were native speakers of English except for one, who was excluded from all analyses. One participant who stopped early was likewise excluded from all analyses.

2.3. Materials

Experimental stimuli included 160 sentence frames in one of three conditions: a high constraint sentence with an expected sentence-final word, a high constraint sentence with an unexpected sentence-final word, and a low constraint sentence with an unexpected sentence-final word. The high-constraint sentences were adapted from those previously used by Federmeier et al. (2007), and from a norming study done by Block and Baldwin (2010). The high-constraint sentences in each set used the same context but differed in their sentence-final words (i.e., a classic 'cloze probability' manipulation, e.g., Wlotko & Federmeier, 2012). The low-constraint sentences in the set used the same unexpected sentence-final word as the high constraint sentences, but differed in the preceding context (i.e., a constraint manipulation, e.g., Federmeier et al., 2007).

An example set is as follows:

(1.1) High-constraint context, expected target word (HighExp): *The prisoners were planning their escape*.
(1.2) High-constraint context, unexpected target word (HighUnexp): *The prisoners were planning their party*.
(1.3) Low-constraint context, unexpected target word (LowUnexp): *Larry chose not to join the party*.

Sentence length was controlled across constraint conditions, with both high-constraint and low constraint sentences having an average length of 10 words. Sentence-final target words were matched across expectedness conditions on a number of lexical features including: word frequency (SUBTLEX$_{us}$ corpus; Brysbaert & New, 2009), number of syllables (English Lexicon Project database; Balota et al., 2007), familiarity and imageability (MRC Psycholinguistics Database; Wilson, 1988), concreteness (Brysbaert, Warriner, & Kuperman, 2014), and emotional valence and arousal (Warriner, Kuperman, & Brysbaert, 2013). Expected sentence-final words had a mean cloze probability of .88 (range = .68–1.00) and unexpected sentence-final words for both high- and low-constraint sentences had a mean cloze probability of .01 (range = .00–.12).

We did not specifically control for the phonological onset similarity of the target words across expectedness conditions. Although this was not critical to our preregistered analyses on amplitude (see below), this was important for our exploratory analyses looking at the onset latency of the N400 effects, as unexpected words with the same phonological onset as an expected target word show longer latency N400 effects than those with differing phonological onsets (see Van Petten, Coulson, Rubin, Plante, & Parks, 1999). Fortunately, a post-hoc analysis revealed that the majority of items did have a different onset, with participants hearing only 6.9% of unexpected words sharing similar onsets to the expected word in high constraint sentences.

Stimuli were recorded by a male native speaker of American English using Adobe Audition software, with an audio sampling rate of 44.1 kHz. The audio was then segmented and trimmed to eliminate silent segments in audio. Sentence-final words and the preceding sentence context were recorded separately and presented in separate audio files to eliminate co-articulation in the sentence-final word and provide a clear word onset, which allows for better time-locking and visualization of auditory sensory ERP components. To create a condition in which there was an induced increase in listening effort, power spectrum matched noise was generated and added to each audio file at 3 dB below the speech signal using Matlab. This SNR was chosen based on prior work showing that this SNR increases listening effort without impairing intelligibility (Payne et al., 2021). Thus, there was both an "in quiet" and "in noise" version of each audio file.

An audibility control task was conducted in the same sample to ensure that the noise levels used for the experimental stimuli were not so high so as to make the speech unintelligible (Payne et al., 2021; Piquado et al., 2012; Tun, O'Kane, & Wingfield, 2002). For these stimuli, the same native speaker of American English was used to record the stimuli. The same procedure was used to create a power spectrum matched masking noise at 3 dB below the speech signal volume. Participants heard three different test sentences (e.g., "*Don't touch the wet paint*") and were tasked with "shadowing" each sentence by repeating out loud each word as it was heard (Marslen-Wilson, 1973). This was done to reduce the contribution of any memory components. Participants showed a 97.94%-word repetition accuracy. It should be noted that the SNR used for our "noise" stimuli was relatively high as compared to the SNRs used by other studies investigating listening effort (e.g., Koelewijn, Zekveld, Festen, Rönnberg, & Kramer, 2012, 2012a; Rogers, 2017; Rogers, Jacoby, & Sommers, 2012; Zekveld et al., 2010). Typically, these types of studies use individualized SNRs that allow participants to recognize 50–84% of the words that they hear. The results from our short shadowing task showed that participants were at or near ceiling in being able to correctly perceive the speech in noise at the SNR that we used. In fact, the SNR for noise used in our study (+3 dB) was at a level that would be commonly experienced in everyday life (Smeds, Wolters, & Rung, 2015; Wu et al., 2018). Therefore, we concluded that any effects seen in subsequent analyses could not be explained by participants lacking the ability to successfully perceive the stimuli. Rather, any effects should be due to increases in listening effort.

To ensure that each stimulus was used in each of the six experimental conditions (each of the three sentence types in both in quiet and in noise), four separate lists were created. To create these lists, we split the 160 sentence frames in half and 80 were used for HighExp sentences and 80 were used for HighUnexp sentences. Because there was no overlap in the sentence contexts and sentence-final words between the HighExp and LowUnexp conditions (see examples above), we used the 80 LowUnexp sentences from the same frames that we used for the 80 HighExp sentences. Thus, there were 80 sentences for each of the constraint and expectancy conditions and there was a total of 240 stimuli per list. Half of the stimuli in each condition were presented with background noise and half with no background noise. Therefore, there were a total of 40 trials per each of the six experimental conditions. See https://osf.io/mfahs/for a supplemental figure detailing the list counterbalancing.

2.4. Procedure

Participants were seated 55 cm from a monitor in a chinrest to stabilize their heads. The ambient lighting level was 140 lux and was selected based on a pilot study that showed this lighting level allowed for most people to have their resting pupil size in the middle of their measured largest and smallest pupil sizes, corresponding to the darkest and brightest ambient lighting settings, respectively. Auditory stimuli were presented binaurally through a Maico MA 41 Audiometer using insert earphones at 65 dB HL. Participants were instructed that they would be required to listen to spoken sentences while fixating on a cross located in the middle of the screen. They were told that they would have a memory test administered to them after the task was over to measure how well they remembered each of the sentences. Participants had self-paced breaks between each trial and were offered breaks throughout the task as needed.

Each of the 240 trials followed the same pattern. After the participant pressed the space bar on the keyboard to end the self-paced break, a white fixation cross appeared in the middle of a black screen for 2000 msec with no audio. This period of time was designed to allow the pupil to adjust to presentation of the cross and for measuring a baseline pupil size. After this baseline period, the experimental auditory sentence began, and the fixation cross remained onscreen. The sentence-final word was presented immediately upon the completion of the context sentence with the fixation cross still present. Finally, a 2000 msec post-stimulus period with no audio was present while the fixation cross remained on the screen. See https://osf.io/3crkd/for a supplemental figure showing a visual representation of a trial.

Immediately upon completing this sentence listening task, participants were given a combined sentence recognition and cued word recall memory test. Participants were visually presented with 120 test sentence frames on a tablet computer, each with the sentence-final word missing. They were instructed to mark whether or not they recognized each sentence as one that they had heard during the experimental task. For the sentences they reported as having heard previously, they were asked to recall the sentence-final word to the best of their ability by typing their response. There was no time limit on the memory test. Sixty of the sentences were ones that they had heard during the task and the other 60 were semantic foils. The 60 sentences that were ones heard previously were taken evenly from each of the six experimental conditions such that there were 10 sentences from each condition. Each semantic foil was created by taking 2 to 4 of the meaning-bearing words from a sentence that the participant had actually heard and were used to create a new semantically similar sentence. For example, if the participant had heard the sentence *Dan recognized John even though he had grown a beard*, the semantic foil they would see in the memory test would be *No one at the reunion recognized Dan because he had grown a* … The inclusion of foils was used to make the recognition task more challenging in order to reduce the likelihood of ceiling performance. This approach has recently been shown to elicit robust effects of listening effort on recognition memory (e.g., Koeritzer, Rogers, Van Engen, & Peelle, 2018; Payne et al., 2021).

2.5. EEG recording and processing

EEG was recorded from 32 evenly spaced silver–silver chloride actiCap slim active electrodes distributed by Brain Products (Brain Vision, LLC, Morrisville, NC, United States of America), following the standard international 10–20 localization system for 32 channels (Jasper, 1958). Electrode impedances were kept below at least 20 kOhms. Electrodes were referenced online to the TP10 electrode and re-referenced offline to the average of the TP10 and TP9 electrodes, which are close to the right and left mastoids, respectively. One electrode was placed beneath the left eye on the infraorbital ridge and was used offline with the Fp1 electrode to create an offline virtual VEOG channel to assist in the detection of eye blinks and vertical eye movements. An offline virtual bipolar HEOG channel was created by taking the difference between the TP10-referenced FT9 and FT10 electrodes to be used for detecting horizontal eye movement artifacts. Continuous EEG was amplified through a BrainAmp DC amplifier and was recorded with a lower cutoff at DC (0 HZ) and an online low pass filter of 1000 Hz at a sampling rate of 500 Hz using BrainVision Recorder software. EEG data were downsampled offline to 250 Hz. Prior to analysis, data were bandpass filtered at .1–30 Hz.

The continuous EEG data were epoched 100 msec before the onset of the sentence-final word and 900 msec after the onset of the sentence-final word. Epoched EEG data were examined for artifacts, including eye blinks, eye movements, flatlines, and signal drifts. Any trials that had been flagged as containing artifacts were excluded from analysis. Thresholds used for each of the artifact detection algorithms were selected for each individual subject through condition-blind visual inspection of the data. Any subjects that had greater than or equal to 40% of their data flagged as containing artifacts were removed from any subsequent analyses. On average, a total of 10.4% ($SD = 8.3\%$; range across participants $\leq$ 1–37.2%) of the trials were flagged as containing artifacts and were excluded from analyses. There were no reliable differences in artifact rates across experimental conditions. See section 2.8 for details on the number of participants excluded from EEG-related analyses.

2.6. Pupillometry recording and processing

Pupil size measurements were continuously recorded from the right eye during each trial using an Eyelink 1000 Plus desktop mounted infrared eye tracker camera distributed by SR Research (SR Research Ltd., Ottawa, ON, Canada). Continuous pupil size measurements were recorded at a rate of 1000 Hz using Eyelink software and were downsampled offline to 50 Hz.

Continuous pupil size data were epoched 200 msec before the onset of the sentence audio until 3000 msec after the onset of sentence audio. This time window allows us to track listening effort, as measured by pupil size, over the course of the sentence from the onset (Zekveld et al., 2018). For exploratory analyses, the data were also epoched from 1000 msec before the onset of the sentence-final word to 0 msec before the onset of the sentence-final word and were baseline corrected using the mean pupil dilation -200-0 msec before the onset of the sentence. This time window allows us to look at pupil size immediately preceding the target word which, if pupil size is an appropriate measure for listening effort, would tell us what kind of effort a listener is experiencing immediately preceding the target word.

Epoched pupil data were examined for artifacts, including eye blinks and pupil dilation speed outliers, which can occur when the camera temporarily detects eyelashes or corrective lenses as part of the pupil and are seen as implausibly fast dilations of the pupil. Once dilation speed outliers were detected, the corresponding data points were removed.[2] Blinks were defined as gaps in the continuous data of more than 75 msec. When a gap this large or larger was detected, 50 msec of data points were removed on either side of the blink. After these first two initial steps, any trial that was missing 40% or more of the data points was flagged for exclusion from subsequent analyses. Using these criteria, only an average of 3.6% of trials were excluded and no participants met the criterion of being excluded from analysis. Any trial that was not excluded had its missing data points filled in by linear interpolation. Next, the interpolated data were run through a 10 Hz low-pass Butterworth filter. Finally, each trial was baseline corrected by dividing each time point by the mean pupil size measured during the 200 msec prior to the onset of the sentence audio. This gave the proportion change from baseline at each time point. This same baseline period was used for both of the epoch periods described above. For analyses that looked at the relationship between pupil mean dilation and other measures, we opted to subject mean-standardize the single trial mean dilations. This allows us to interpret trial-to-trial change as a function of an individual's own average, thus removing between-subjects variation from within-person analyses (Enders & Tofighi, 2007).

2.7. *Electrophysiological data analyses*

Thirty-nine of the available 42 subjects were used for ERP analyses. Two subjects were dropped because they had more than 40% of their data flagged as containing artifacts. An additional subject was dropped because over 40% of their EEG data were missing due to experimenter error.

Planned analyses of the N400 amplitude response to sentence-final words were conducted using linear mixed-effects models. Fixed-effects for noise, target-word type, and their interaction were used. Random-effects structures were defined to represent the experimental design and nested sampling structure seen in our data. Therefore, we used random intercept terms for subject and electrode and random slopes across subjects for noise, target word type, and their interaction. These models were fit using the *lme4* package in the R statistical software (Bates, Mächler, Bolker, & Walker, 2014). N400 analyses were conducted across six centro-parietal electrode sites (CP1, CP2, Cz, P3, P4, and Pz), where the N400 effects are typically largest. Mean amplitudes were calculated within the 300–500 msec time window, which was selected a priori. For this and all subsequent analyses using mixed effects models, statistical inference on the fixed effects was done using separate likelihood ratio tests for each of the fixed-effect parameters. Likelihood ratio tests were computed using the *mixed()* function from the *afex* package in R (Singmann et al., 2015). For this and subsequent analyses, for follow-up tests decomposing higher-order interactions, we calculated pairwise contrasts on the estimated marginal means (sometimes called least-squares means) calculated using the *emmeans()* function from the *emmeans* package in R (Lenth, Singmann, Love, Buerkner, & Herve, 2019). Adjustments for multiple comparisons on all analyses were done using the false discovery rate procedure. We predicted that if the use of context is inhibited when listening to degraded speech, we should see a reduction in the N400 expectancy effect. But if context use is differentially relied upon, then there may be an associated increase in the N400 expectancy effect to speech in noise.

In an exploratory analysis, we tested whether listening in noise had any effect on the onset latency of the N400 effect. Difference waves of the expectancy effect were constructed via pointwise subtraction of the subject ERP waveforms for the HighUnexp and HighExp conditions separately for the noise and quiet conditions. Using these difference waves, raster plots were created by calculating false discovery rate corrected *t*-statistics at each time point and plotting these separately for the quiet and noise conditions. If the raster plot contained any significant differences within the 200–600 msec time window, we proceeded to use a jackknife-based procedure to measure onset latency (Kiesel, Miller, Jolicœur, & Brisson, 2008; Ulrich & Miller, 2001). This larger time window was used because there has been evidence that N400-like activity for auditory stimuli may start to emerge as early as 200 msec after onset (Van Petten et al., 1999). To do this, the 50% peak latency was calculated for each jackknife subsample from the Cz electrode using the subject-level difference waves. A jackknife-corrected *t* test (see Kiesel et al., 2008; Ulrich & Miller, 2001) was conducted to compare the onset latencies of the N400 effect difference wave between the noise condition and the quiet condition. A similar procedure was done to inspect the constraint effect, which can be seen by looking at the HighUnexp – LowUnexp difference waves.

[2] To detect speed dilation outliers first a vector of dilation speeds is created, called d'. This vector is created using the following formula: $d'[i] = \max\left(\left|\frac{d[i]-d[i-1]}{t[i]-t[i-1]}\right|, \left|\frac{d[i+1]-d[i]}{t[i+1]-t[i]}\right|\right)$ Where d' is the pupil size as point *I* and *t*[i] is the time at point *I* (in msec or whatever scale is being used). Next, the median absolute deviation (MAD) is calculated: $MAD = median(|d'[i] - median(d')|)$ After calculating MAD a cutoff threshold is calculated: $Threshold = median(d') + n{*}MAD$ Finally, once the threshold is set compare each value in d' with the threshold and any point that exceeds this threshold will have its data point removed. This was adopted from Kret and Sjak-Shie (2019).

2.8. Behavioral data analyses

The following analyses of the memory test data were confirmatory. Of the 42 participants available for analysis, only 38 were used for memory test analyses. Two participants experienced technical errors while taking the test and their data were unusable, one of the participants was administered the wrong memory test, and one asked to end the study early after only completing 17 of the 120 test questions. Analyses were conducted separately for the recognition memory and cued recall portions of the test. For the behavioral analyses, linear mixed effects models were fit with random-effects structures (described below) that were kept maximal enough to allow for convergence and avoid singular fit (see Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015).

For the recognition portion of the memory test, hit rate scores were aggregated for each participant and for each of the experimental conditions. We calculated scores for noise and quiet conditions for both high-constraint sentences and low-constraint sentences and thus, ended up with four scores per subject. For the linear mixed effects model, hit rate was modeled as a function of noise, contextual constraint, and their interaction as predictor variables. A random intercept for subject was used with no random slopes as these did not allow for convergence (see e.g., Bates et al., 2015).

Analysis of the recall portion of the memory test followed a procedure that was similar to what was done for the recognition data. First, we aggregated the data to calculate the proportion correct in each of the relevant experimental conditions for each participant. For the recall data this was done for both the noise and the quiet conditions for expected and unexpected sentence-final words heard in high-constraint and for sentence-final words heard in low-constraint sentences. Thus, there were six separate recall scores calculated for each participant. A linear mixed effects model was fit following the same procedure used for the recognition memory data. Fixed-effects for noise, sentence-final word type, and their interaction were used. We used a random-effects structure using a random intercept for subject (random slopes were not used as they did not allow for convergence or avoidance of singular fit).

We predicted a priori that recognition hit rate and recall accuracy would generally be lower for sentences and words heard in the presence of background noise. If the use of context is hindered when listening to degraded speech, as compared to speech in quiet, the effect of noise on memory measures should lead to there being less of a difference between recognition performance for high constraint and low constraint sentences in the noise condition. In contrast, if listeners differentially rely on context when listening to speech in noise, this would be reflected in a reduction of the negative effects of background noise on memory measures in highly constraining contexts (e.g., McCoy et al., 2005).

2.9. Pupillometry data analyses

According to the criteria described above no participants needed to be dropped from analysis of the pupillometry data. Mean proportion change in pupil size from baseline was calculated -1000 - 0 msec prior to the onset of the sentence-final word.[3] Therefore, we calculated the mean proportion change for each subject for both the noise and the quiet experimental conditions, collapsing across the contextual constraint experimental manipulations.[4]

According to our preregistered analysis plan, a linear mixed-effects model was fit using mean proportion change from baseline of pupil size as the response variable and noise as the predictor variable. A random intercept for subject was fit, which was the maximal random effects structure that allowed for convergence. Before collecting the data, we hypothesized that we would see patterns similar to those seen previously (for a review see Zekveld et al., 2018), such that the mean dilation of the pupillary response would be larger when listening in noise than when listening in quiet. We predicted that if this was the case, then the pupillary response would be a valid measure of listening effort.

2.10. ERP-pupillometry coupling data analyses

The subjects used for these analyses were the same 39 used for the ERP-only analyses described above. All analyses of the relationship between the pupillary response and other outcomes were conducted in the noise condition only, reasoning that variation in pupil dilation during the processing of acoustically challenging speech would reflect variation in listening effort. These ERP-pupillometry coupling analyses were conducted on the single-trial mean amplitudes for both the ERP and pupil data.

First, in a planned analysis, we tested the relationship between pupil size and the N400 amplitude response while listening to speech in noise. We measured N400 single-trial mean amplitudes from the same measurement window and electrodes as the averaged ERP analyses. We measured single-trial mean proportion change from baseline in pupil size from the -1000 – 0 msec time window as described above. For these, and subsequent analyses comparing pupil size changes with other measures, we used the time window for pupil size 1000 msec prior to the onset of the sentence final word. Mean proportion change in pupil diameter from baseline was calculated using a 200 msec baseline time period prior to the onset of the sentence. These scores were then subject-mean standardized as in the grand-average pupil analyses described above. Following this, a linear mixed effects model

[3] We preregistered using the pupillary response time-locked to the onset of the sentence rather than the onset of the sentence-final word. We did run this analysis and it had an almost identical outcome as using the mean dilation time-locked to the onset of the sentence-final word. Therefore, in this document we decided to use the later to remain consistent with the subsequent coupling analyses. This supplementary analysis can be found in the document at https://osf.io/3u65g/.

[4] The preregistration for this analysis reported that we would also look at context effects as well. However, the main purpose of this specific analysis was to test whether pupil size could be used as a reliable measure of noise-induced listening effort, rather than examining context effects on pupil size. Indeed, upon further reflection, it is implausible to expect any constraint effects at the onset of the sentences. However, because we preregistered this analysis we did run it and unsurprisingly found no significant effects of context on pupil size during this time window.

was fit with N400 single-trial mean amplitude (averaged across all electrodes) as the response variable. Target word type, pupil size, and their interaction were used as predictor variables, in order to test whether trial-to-trial variation in pupil size predicts N400 mean amplitude. The maximal random-effects structure that would allow for convergence included a random intercept for subject (i.e., including random slopes would not allow for convergence). The *emtrends()* function from the *emmeans* package in R was used explore significant interaction by calculating simple slopes and their corresponding 95% confidence intervals, as well as testing for significant differences between the simple slopes (Aiken, West, & Reno, 1991). Our preregistered hypotheses were as follows: if the use of context is inhibited by increases in listening effort, then we should see that on trials with a larger pupillary response, there would be a reduced N400 effect. However, if the listener becomes more reliant on context as listening effort increases, then we should see that the N400 effect increases with increasing pupil dilation.

To understand how changes in listening effort may affect the onset of the N400 effect, an exploratory analysis looking at the relationship between the onset latency of the N400 effect and pupil size while listening in noise was also conducted. To do this, we first aligned single-trial epoched EEG data with single-trial mean amplitude pupil dilation data. As above, we used only those trials in which a participant was listening to speech in noise. We next split the epoched EEG data into two sets, with one set containing trials that had mean pupil dilations that were lower than the median pupil size for a subject and the other set containing trials that were greater than or equal to the median pupil size for a subject. Thus, we binned trials together that had a smaller pupillary response for that subject and those with a larger pupillary response for that subject when listening to speech in noise. We used these binned trials to create difference waves of the expectancy effect (HU - HE) separately for the trials with smaller pupil sizes and for the trials with larger pupil sizes. Using these difference waves, raster plots were then created by calculating false discovery rate corrected *t*-statistics at each time point and plotting these separately for the small and large pupil size trials. We then used the jackknife grand-average method (Kiesel et al., 2008; Ulrich & Miller, 2001) to find the 50% peak latency onset of the expectancy effect separately for large pupil trials and small pupil trials. A jackknife-corrected *t* test was calculated between the onset latencies to test if there were differences in the onset of the N400 effect as a function of listening effort.

2.11. Behavioral-pupillometry coupling data analyses

The analyses described in this section were exploratory tests of whether trial-to-trial variation in pupil size when listening to speech in noise predicts subsequent memory performance. The same subjects that were used for the behavioral-only analyses described above were used for these analyses.

For the measure of recognition memory from the memory test, we fit a generalized linear mixed effects model to the single trials, assuming a binomial distribution using a logit link function, using the *glmer()* function from the *lme4* package in R (Bates et al., 2014). Single-trial recognition memory accuracy was the dependent variable. We used sentence-level contextual constraint (high *vs* low), subject-standardized mean proportion change in pupil size, and their interaction as predictor variables. A random intercept for subject and a random slope for context was used, which was the maximal random effects structure that allowed for convergence. Odds ratios were inspected to interpret the effect size magnitude of any significant effects.

For the recall measure from the memory test, we fit a generalized mixed-effects model that was similar to the one we fit for the recognition data. The response variable was single-trial accuracy for recalling the sentence-final word and target-word type, pupil size, and their interaction were the predictor variables. A random intercept for subjects was fit, which was the maximal random effects structure allowing for convergence of the model.

2.12. Frontal positivity analyses

Analyses to explore the late frontal positivity that was first reported by Federmeier et al. (2007) were also pre-registered. Typically, this positivity is seen to words that are unexpected in a highly constraining contexts and is thought to reflect the response to having strong predictions violated. This positivity has mostly been seen over prefrontal and frontal electrodes and begins immediately after the N400 (Federmeier, Kutas, & Schul, 2010). However, our data provided no evidence of a frontal positivity-like effect (see Fig. 1), and analyses examining effects of noise and listening effort (via pupil dilation) produced no significant findings (see https://osf.io/2wrz7/). As such, we do not further discuss the frontal positivity results below.

3. Results

3.1. Pre-registered N400 mean amplitude analyses

The grand average ERPs for the average across all posterior electrodes can be seen in Fig. 1. Note, the gray regions indicate the time windows used for calculating the mean amplitudes used in ERP mean amplitude analyses. For the N400 amplitude analyses, we found that there was a significant main effect of sentence-final word type (χ^2 (2) = 41.43, $p < .01$) but no main effect of noise (χ^2 (1) = .19, $p = .67$). However, these effects were qualified by a significant interaction between word type and noise (χ^2 (2) = 6.09, $p < .05$). We explored this interaction by calculating contrasts between the estimated marginal means. The results of these pairwise comparisons can be found in the top panel of Table 1. These post-hoc contrasts revealed a classic N400 pattern, with a larger N400 mean amplitude for unexpected words compared to expected words. The mean amplitudes for HighExp and LowUnexp were unaffected by the noise manipulations. However, we found a reduction of the mean amplitude in noise for HighUnexp that was marginally significant. Importantly, we found that while the mean N400 amplitude was larger for HighUnexp than for HighExp both in quiet and in noise, this effect was significantly reduced in magnitude in noise compared to quiet.

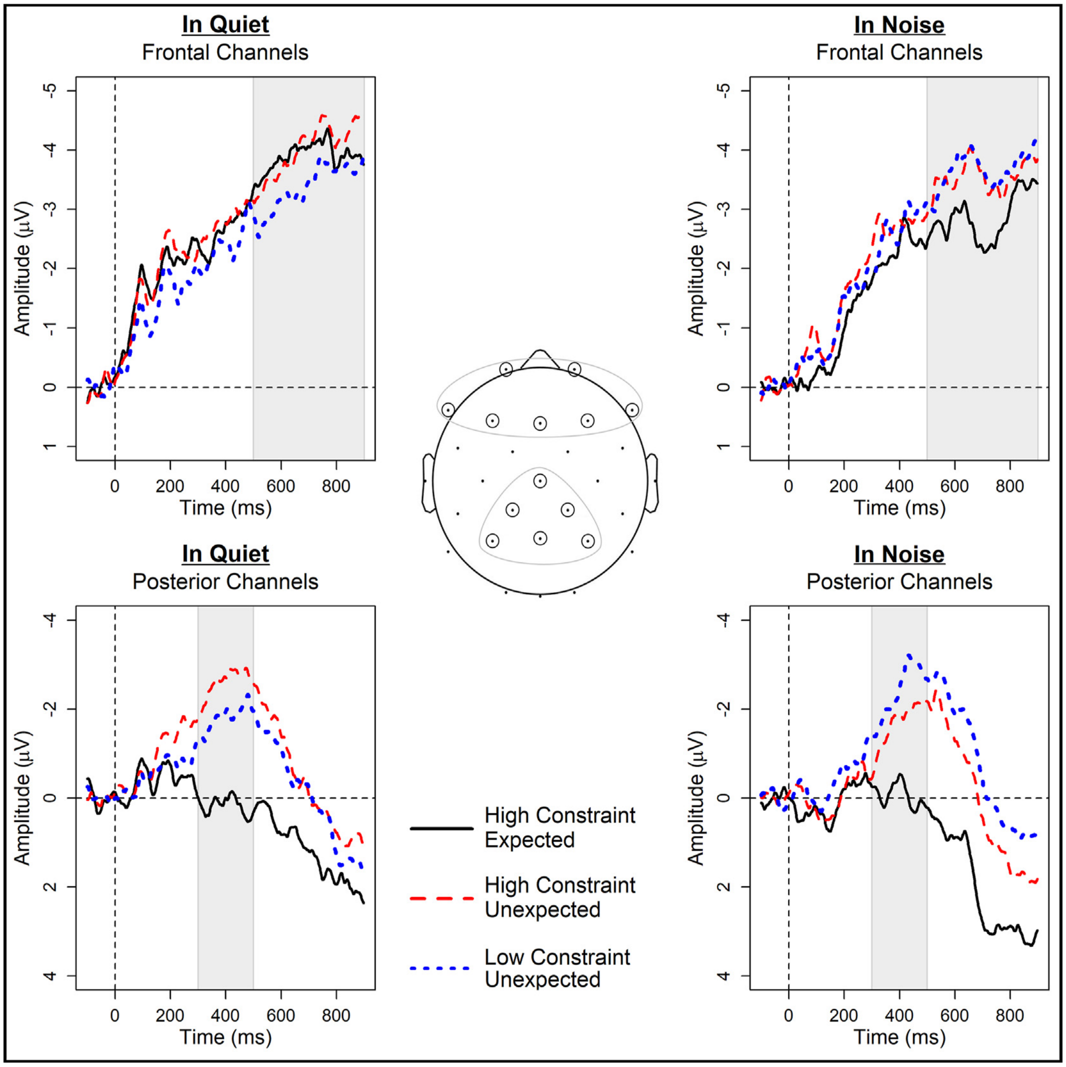

Fig. 1 – ERPs as a function of noise and sentence type. The average of the frontal and posterior channels used in the analyses on the mean amplitudes are plotted (grouped channels are indicated on the scalp map in the middle). Grayed areas show the time windows used for calculating mean amplitudes. ERPs on the left are for sentence-final words heard without any background noise present, while ERPs on the right are for those heard with background noise.

3.2. *Exploratory N400 latency analyses*

Fig. 2A shows the ERP difference wave for the expectancy effect (HighUnexp – HighExp) in quiet and in noise for the posterior electrodes. Fig. 2B shows the scalp distribution of these effects over time and Fig. 2C shows FDR corrected raster plots of the expectancy effect at each electrode site. As Fig. 2C shows, there were clear differences in the onset of the N400 expectancy effect in noise compared to quiet. Note from Fig. 2B and C that the expectancy effect shows a canonical N400 centro-posterior distribution. Using the jackknife latency analysis described above, we found that onset latency when listening in quiet was 299.85 msec and was 372.92 msec when listening in noise (a difference of 73.08 msec). This difference in latency onset was statistically significant ($t\_corrected\ (38) = -3.20$, $p < .01$). Similar raster plots for the constraint effect (HighUnexp – LowUnexp) showed that this contrast was not significant at any time point for any electrode (see https://osf.io/3u65g/). Therefore, we did not pursue any latency analysis for this effect.

3.3. *Pre-registered memory analyses*

The left-most panel of Fig. 3A shows the estimated marginal means of recognition hit rate.[5] For the recognition memory analysis, we found that there was a significant main effect of

[5] To check for response bias, we ran an analysis comparing criterion location between conditions (c; a measure of response bias). We found that low constraint sentences had a significantly higher c value than high constraint values. We found a small but significant bias ($c = .39$) for participants to respond “no, I do not remember this sentence” to low constraint sentences. However, the bias for high constraint sentences was not significantly different from 0, indicating no bias. Therefore, it is highly unlikely that response bias is accounting for the observed hit rate patterns seen here.

Table 1 – Pairwise post-hoc contrasts for three of the univariate models. The top panel reports the N400 response. The middle panel reports recognition memory hit rate. The bottom panel reports recall accuracy. HighExp = High constraint sentence with expected sentence-final word; HighUnexp = High constraint sentence with unexpected sentence-final word; LowUnexp = Low constraint sentence with unexpected sentence-final word.

N400 Mean Amplitude Model Post-hoc Contrasts				
Contrast	Dif. Est.	t (df = 40)	p-value	95% CI
Quiet: HighExp versus HighUnexp	2.76	6.35	<.01	[1.46, 4.06]
Noise: HighExp versus HighUnexp	1.64	3.55	<.01	[.26, 3.01]
Quiet: HighExp versus LowUnexp	1.75	3.59	<.01	[.29, 3.21]
Noise: HighExp versus LowUnexp	2.07	4.65	<.01	[.74, 3.41]
Quiet: HighUnexp versus LowUnexp	-1.01	-1.69	.15	[-2.81, .79]
Noise: HighUnexp versus LowUnexp	.44	.97	.39	[-.92, 1.80]
HighExp: Quiet versus Noise	-.16	-.35	.90	[-1.49, 1.18]
HighUnexp: Quiet versus Noise	.97	2.21	.05	[-.34, 2.28]
LowUnexp: Quiet versus Noise	.48	-1.10	.38	[-1.79, .83]
Recognition Memory Model Post-hoc Contrasts				
Contrast	Dif. Est.	t(df = 117)	p-value	95% CI
High Constraint: Quiet versus Noise	-.01	-.20	.84	[-.07, .06]
Low Constraint: Quiet versus Noise	.08	2.60	<.05	[.02, .15]
Quiet: High versus Low Constraint	.22	7.01	<.01	[.16, .29]
Noise: High versus Low Constraint	.31	9.81	<.01	[.25, .37]
Recall Memory Model Post-hoc Contrasts				
Contrast	Dif. Est.	t(df = 195)	p-value	95% CI
HighExp versus HighUnexp	.34	13.51	<.01	[.28, .40]
HighExp versus LowUnexp	.48	19.01	<.01	[.42, .54]
HighUnexp versus LowUnexp	.14	5.50	<.01	[.08, .20]
Quiet versus Noise	.05	2.35	<.05	[.01,0.09]

contextual constraint (χ^2 (1) = 93.62, $p < .01$) but there was not a significant main effect of noise (χ^2 (1) = 2.93, $p = .09$).

However, there was a significant interaction between constraint and noise (χ^2 (1) = 3.96, $p < .05$). The results of pairwise contrasts can be found in the middle panel of Table 1. These post-hoc comparisons showed that higher constraint sentences were remembered better than lower constraint sentences both in quiet and in noise. We found that for low constraint sentences, recognition was significantly worse in noise compared to quiet. However, for higher constraint sentences, this noise effect was reduced to non-significance.

The results from the recall portion of the memory test can be seen in the right-most panel of Fig. 3A. The bottom panel of Table 1 shows the pairwise contrasts used for post-hoc comparisons on the model fit for recall accuracy. For the recall portion of the memory test, we found a main effect of target word type (χ^2 (2) = 213.07, $p < .01$). Contrasts showed that HighExp words were remembered better than both HighUnexp and LowUnexp words. Moreover, there was a main effect of background noise (χ^2 (1) = 5.60, $p < .05$), such that sentence-final words heard in quiet were remembered significantly better than those heard in noise. However, we did not find a significant interaction between these effects (χ^2 (2) = .52, $p = .77$).

3.4. *Pre-registered pupillometry analysis*

Fig. 3B shows the task evoked pupillary response across the duration of listening to a sentence as well as the change in pupil size relative to baseline just prior to the onset of the sentence-final word. Note that Fig. 3B shows that about 500 msec after the onset of the sentence, the pupil size starts to increase relative to baseline, with a larger pupillary response for sentences heard in noise. This pattern is present for the pupillary response time-locked to the onset of the sentence-final word as well. Therefore, to remain consistent with subsequent analyses, we used the mean dilation for the 1000 msec prior to the onset of the sentence-final word. We found that there was a significant effect of noise on the pupillary response (χ^2 (1) = 9.59, $p < .01$), such that there was a larger pupil size when listening in noise versus when listening in quiet (t (43) = 3.24, $p < .01$).

3.5. *Pre-registered ERP-Pupillometry coupling analyses*

Fig. 4B shows raster plots of the expectancy effect (HU-HE) separately for trials with larger (above the intra-subject median) versus smaller (below the intra-subject median) pupillary responses and Fig. 4A shows the results from our analyses of the single-trial relationship between mean N400 amplitude and pupillary responses. We found a main effect of sentence-final word type (χ^2 (2) = 33.47, $p < .01$) but no main effect of pupil size (χ^2 (1) = .30, $p = .58$) on single trial N400 amplitude. However, there was a significant interaction between target word type and pupil size (χ^2 (2) = 12.60, $p < .01$). Table 2 contains the results on simple slope estimates and pairwise comparisons of these simple slopes. Simple slopes estimates indicated that LowUnexp N400 mean amplitude was unaffected by the mean pupil response. However, HighExp and HighUnexp N400 amplitudes both had a significant, but opposite, relationship with pupil size. The N400 response to HighExp decreased with increasing pupil size, while this response increased to HighUnexp with increasing pupil size. Importantly, these findings indicate that the expectancy effect (or the difference between the HighUnexp and HighExp N400 responses) *increased* with increases in listening effort, as measured by the pupil response.

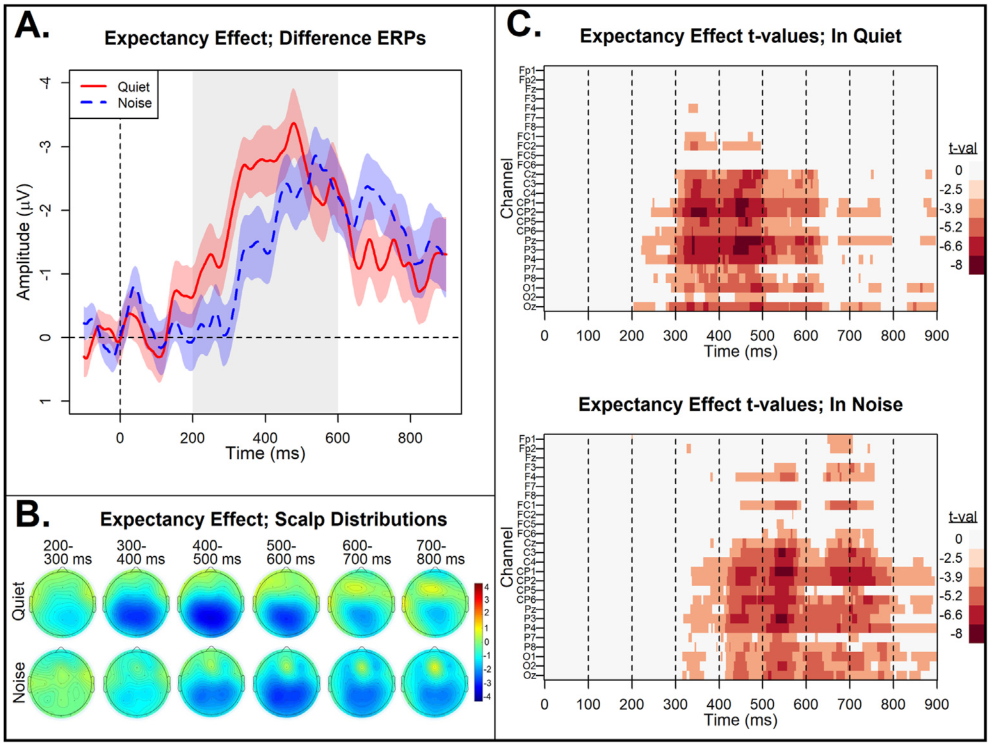

Fig. 2 – The N400 expectancy effect as a function of noise. For each plot, the expectancy effect was calculated by taking the pointwise difference between the unexpected words and expected words (i.e., HighUnexp – HighExp) in the high constraint sentences. A. Difference wave ERPs for sentences heard in quiet (solid, red) and in noise (dashed, blue). Shaded red and blue areas show the standard error of the mean at each timepoint. The gray shaded area shows the time window used (200–600 msec) in the onset latency jackknife analyses. B. Scalp topography maps (in 100 msec bins) from 200 to 800 msec highlight the latency shift of the N400 expectancy effect when listening in noise. The top row is the quiet condition and the bottom row is the noise condition. C. False discovery rate corrected raster plot of the expectancy effect (see text for detail). Channel is depicted on the Y-axis - anterior electrodes appear in the top rows, down to the posterior electrodes in the bottom rows.

3.6. *Exploratory ERP-Pupillometry latency coupling analyses*

Onset latency analysis using jackknifed subsamples found that there was no significant difference between trials with large pupillary responses and trials with small pupillary responses ($t\_corrected\ (38) = -.46$, $p = .65$). This could be due to the fact that there is almost no expectancy effect for trials which had a smaller pupillary response (see Fig. 3B). If there is no peak to use for onset latency analysis, then the calculated onset latencies will have a large error variance (Kiesel et al., 2008; Ulrich & Miller, 2001). Indeed, we found that the onset latency for trials in which there was a larger pupillary response had a stable onset latency, with a calculated average latency of 331.49 msec and a SE of .74 (across jackknife subsamples). However, for trials with a smaller pupillary response, there was a much less stable onset latency, with an average of 380.72 msec and a SE of 2.72. Therefore, in line with the raster plots and single trial analysis, there did not appear to be a reliable N400 expectancy effect among trials with lower pupil dilation responses to noise.

3.7. *Exploratory behavioral-pupillometry coupling analyses*

Fig. 4C displays the pupillary response 1000 msec prior to the onset of the sentence-final word. These pupillary responses were binned based on subsequent memory performance. We observed a larger pupillary response during speech processing for sentences or words that were subsequently forgotten compared to those that were remembered. The results from our analyses confirmed this. For the recognition memory data, we found a main effect of constraint (χ^2 (1) = 40.08, $p < .01$) such that the odds of recognizing a highly constraining sentence was 4.42 times the odds of recognizing a low constraint sentence. Importantly, we also found a significant effect of pupil size (χ^2 (1) = 4.10, $p < .05$) such that the odds of recognizing a sentence increased 1.29 times with each standard deviation *decrease* in pupil dilation. There was no significant interaction between these effects (χ^2 (1) = 2.01, $p = .16$).

The results from the recall portion of the memory test were similar. We found that there was a main effect of target word type (χ^2 (2) = 196.30, $p < .01$) with the odds of recalling an

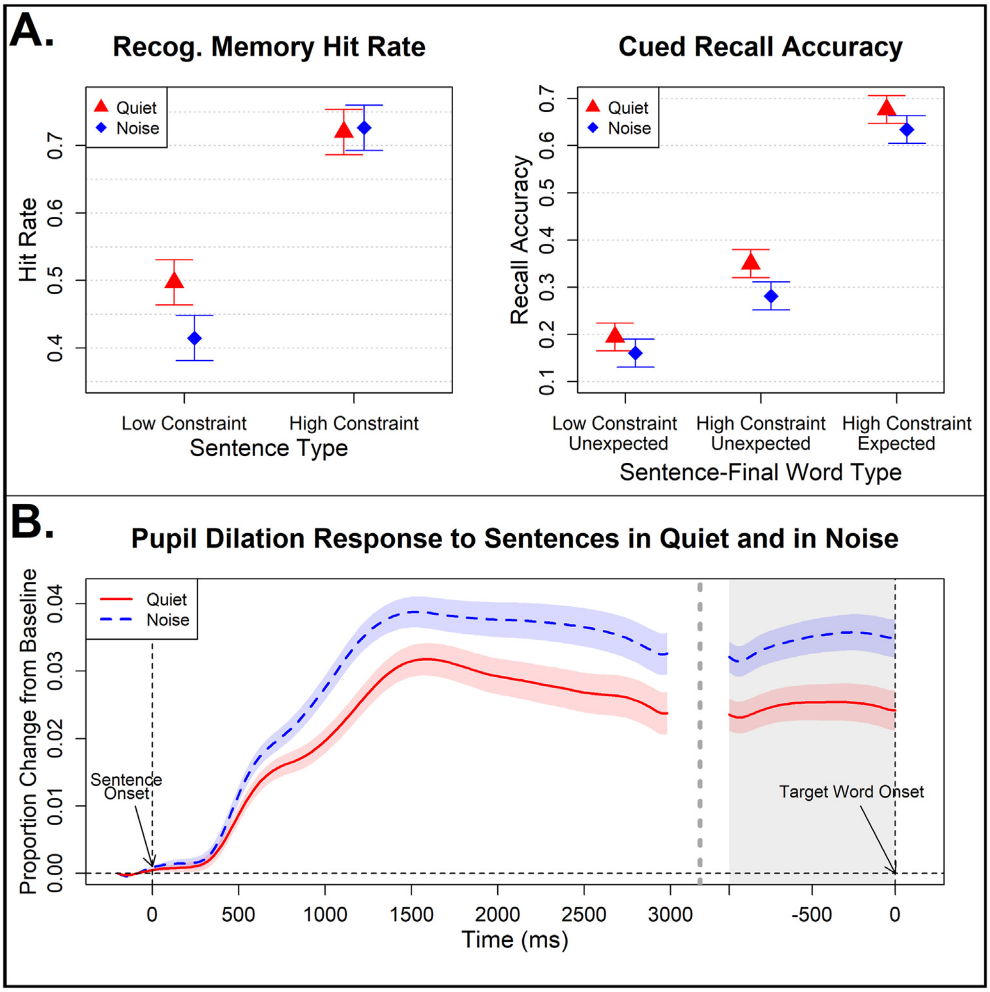

Fig. 3 – Behavioral and pupillary responses. A. Results for the memory test. The left panel shows the hit rate on the sentence recognition portion of the test as a function of noise and sentence type. The right panel shows mean accuracy on the cued word recall portion of the test as a function of noise and sentence-final word type. Error bars represent the estimated standard error of the mean. B. Proportion change in pupil size from baseline. Left of the dotted gray line shows the pupillary dilation response time locked to the onset of the sentence. On the right, the pupillary response is plotted time locked to the onset of the sentence-final word.

expected word in highly constraining context being 4.65 times the odds of recalling an unexpected word in highly constraining context and 10.07 times the odds of recalling an unexpected word heard in low contextual constraint. There was also a significant effect of pupil dilation (χ^2 (1) = 5.68, $p < .05$), with the odds of recalling a word correctly increasing by 1.12 times with each standard deviation *decrease* in pupil dilation. There was no significant interaction between target word type and pupil dilation (χ^2 (2) = 2.53, $p = .28$).

4. Discussion

In this study, we examined how the amount of effort experienced by a listener affects how they use contextual information when listening to speech. According to models of listening effort (Peelle, 2018; Pichora-Fuller et al., 2016), even small amounts of background noise should result in increased effort at perceptual decoding, negatively impacting higher-level online and offline speech processing, even when a listener can perceive the speech correctly (McCoy et al., 2005; Payne et al., 2021; Piquado et al., 2012; Rabbitt, 1968, 1991). Typically, those within the field of cognitive audiology theorize that contextual information can be supportive, helping the listener to overcome the negative effects of listening effort (e.g., Pichora-Fuller, 2008). In contrast, the field of cognitive electrophysiology has generally theorized that the use of context (as measured by the N400) is impaired when listening to perceptually challenging speech (e.g., Obleser & Kotz, 2011; Aydelott et al., 2006; Strauß et al., 2013).

We found that when using methodologies from both of these fields that there was evidence supporting both of these apparently contrasting hypotheses. When viewing the recognition memory results *in isolation*, we found evidence that contextual information helps to reduce negative effects of noise. But when viewing the N400 results in isolation, it appears that context use is impaired when listening in noise. Taken together these contrasting results within the same

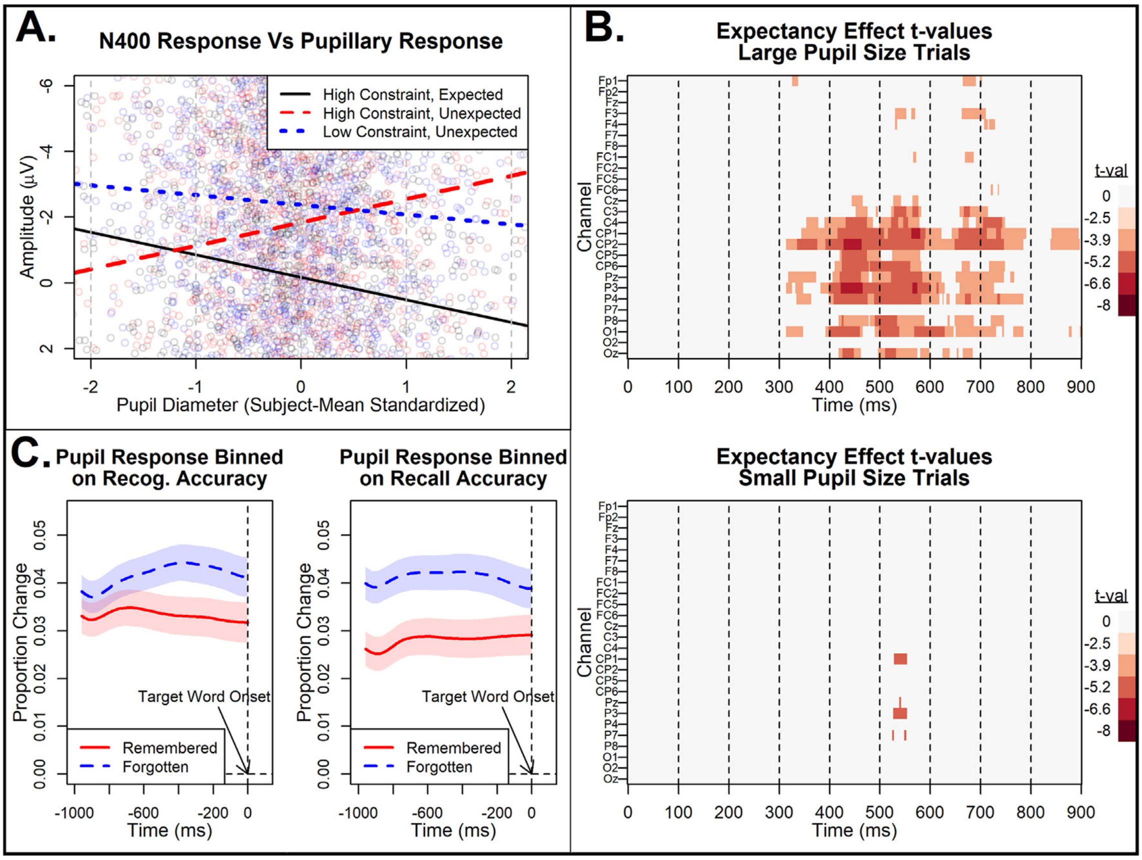

Fig. 4 – The relationship between the pupillary response and N400 and memory outcomes when listening to speech in noise. A. Scatter plot showing the relationship between single-trial N400 mean amplitude and subject-mean standardized pupillary response. B. FDR-corrected raster plots of the expectancy effect as a function of pupil size. Top raster plot shows results from trials with larger pupillary responses (above the intra-subject median) while the bottom raster plot shows the results for smaller pupillary responses (below the intra-subject median). C. The pupillary response time-locked prior to the onset of the sentence-final word as a function of subsequent memory performance. Left plot: Pupillary response based on subsequent recognition memory performance. Right plot: Pupillary response based on subsequent recall performance.

Table 2 – Simple slope estimates and contrasts for the N400 mean amplitude and pupillometry coupling analysis. HighExp = High constraint sentence with expected sentence-final word; HighUnexp = High constraint sentence with unexpected sentence-final word; LowUnexp = Low constraint sentence with unexpected sentence-final word.

Simple Slope Estimates		
Effect	Estimate	95% CI
High Constraint, Expected	.69	[.13, 1.24]
High Constraint, Unexpected	-1.27	[-1.27, -.15]
Low Constraint, Unexpected	.30	[-.28, .88]

Simple Slope Contrasts				
Contrast	Dif. Est.	z	p-value	95% CI
HighExp versus HighUnexp	1.40	3.45	<.01	[.43, 2.36]
HighExp versus LowUnexp	.39	.95	.34	[-.60, 1.37]
HighUnexp versus LowUnexp	-1.01	-2.45	<.05	[-1.99, -.02]

participants paint a more dynamic picture of context use when listening to degraded speech. That is, contextual information may not be able to be used as efficiently for the processing of individual words, but it may still be used just as effectively to help construct and maintain a "good enough" sentence-level representation (Ferreira et al., 2002; Ferreira and Patson, 2007; Ferreira and Lowder, 2016). Importantly however, this set of results represents how the use of context changes when listening in noise *generally* (i.e., effects of acoustic challenge). When we looked more directly at the effects of listening effort, as reflected in the pupillary response on ERP and memory outcomes, we found that context use varies dynamically from sentence to sentence depending on how the listener *responds* to perceptual challenge. When a listener exerted greater effort, as reflected by an increased pupil dilation response when listening to degraded speech, they were able to recover the use of contextual information for online processing of speech, as measured by the N400. However, this increase in effort was also accompanied by a general reduction in memory, suggesting an effort-driven resource trade-off between word processing and subsequent memory, consistent with the predictions of listening effort theories, such as FUEL (e.g., Rabbitt, 1968, 1991; Pichora-Fuller et al., 2016; Peelle, 2018; Zekveld et al., 2018). In the following sections, we discuss these findings in more detail and discuss their implications for theories of context processing and listening effort in speech perception.

4.1. *Effects of acoustic challenge on speech memory*

We found, overall, that sentences that had highly constraining contexts were recognized much better than low constraint sentences. Additionally, expected words heard in highly constraining contexts were recalled with much higher accuracy than unexpected words. This suggests that a supportive semantic context helps a listener build better long-term memory representations. Perhaps this is because predictive processes help to alleviate some of the burden of processing, allowing for more resources to be available for memory encoding and maintenance. In fact, we did find some evidence that might suggest that prediction violations (unexpected words in highly constraining contexts) are remembered better than unexpected words in low constraint sentences. This suggests that prediction violations may be encoded more deeply than unexpected words that are not embedded within supportive contexts that afford predictive processes (see also Ferreira & Lowder, 2016; Rommers & Federmeier, 2018). Therefore, while sentential constraint overall seems to provide a benefit to recognition by potentially allowing for more resources to be available for encoding, maintenance and/or retrieval, when an encountered word violates a strongly held prediction, it may be processed more deeply than when that word is encountered in situations that do not afford strong predictions. Alternatively, these particular data are also consistent with a bottom-up account that does not require prediction. Highly constraining contexts may afford less effortful construction of sentence-level semantic representations from the bottom-up signal. This, in turn, could allow for deeper encoding of memory representations, making it easier for later recollection.

We observed that both delayed recognition memory hit rate for low constraint sentences and general recall accuracy were negatively impacted by listening to speech that was accompanied by background noise. These effects were present even though participants' accuracy on the shadowing task was near ceiling, showing that they could correctly perceive the speech at the SNR used for the main task. This finding replicates past work showing that acoustic challenge interferes with memory processes (Cousins, Dar, Wingfield, & Miller, 2014; Koeritzer et al., 2018; McCoy et al., 2005; Piquado et al., 2012; Rabbitt, 1968, 1991; Van Engen, Chandrasekaran, & Smiljanic, 2012; Wingfield, Tun, & McCoy, 2005; Payne et al., 2021). At the same time, we found that highly constraining sentential context seemed to completely eliminate the negative effects of noise on recognition memory that was observed for low constraint sentences. This is in agreement with a number of prior studies that have found that supportive context leads to better memory performance (for a review see Payne & Silcox, 2019).

Interestingly, we did not find evidence for the same compensatory effects in word recall. Although we saw that expected words were generally remembered better than unexpected words, we found that listening in noise decreased performance on the recall portion of the memory test similarly for all types of sentence-final words. This suggests that there was a dissociation in the beneficial effects of context for sentence recognition and word recall, with only sentence recognition in noise showing selective improvement with increasing constraint. Decades of work establishing functional differences between cued recall and recognition memory (e.g., Craik & McDowd, 1987; Danckert & Craik & Lockhart, 1972; Jacoby, Toth, & Yonelinas, 1993, 1979; Mandler, 1980; Rugg & Yonelinas, 2003; Yonelinas, 2002) have shown that successful recall relies primarily on more effortful recollection processes, whereas recognition memory can be supported in part by weaker familiarity signals. Therefore, it is possible in our study that the observed differences in the effects of context on recall versus recognition of speech in noise are driven by a differential benefit of context on familiarity, benefiting recognition. On the other hand, it could be that more effortful and explicit recollection was not differentially improved by more constraining sentential contexts resulting in a reduced effect on recall. Additionally, the observed differences could possibly be driven by differences in the recall task focusing on word-level recall, whereas the recognition task focused on sentence-level retrieval. This is consistent with hierarchical models of verbal memory (Craik, 2002; Ferreira & Patson, 2007; Kintsch, 1998; Kintsch & Mangalath, 2011) which suggest gist-based sentence-level representations are distinct from surface-level lexical representations, which are less well represented in long-term memory. Under this account, supportive contexts may be more beneficial to sentence-level representations in helping to buffer against the negative influences of degraded speech but may be less beneficial to fleeting surface lexical representations.

4.2. *Acoustic challenge and electrophysiological responses*

We observed the typically seen N400 response to expectancy and constraint. Replicating decades of prior research (see Kutas & Federmeier, 2011), we found that N400 amplitude was reduced to expected words and larger to unexpected words. Moreover, N400 amplitude for unexpected words did not significantly differ as a function of contextual constraint. This type of pattern mirrors what Federmeier et al. (2007) observed, when using similar stimuli presented visually (see also, Ng, Payne, Steen, Stine-Morrow, & Federmeier, 2017; Payne & Federmeier, 2017a, 2019; Wlotko & Federmeier, 2007, 2012). Importantly, we found that the amplitude of the N400 expectancy effect (i.e., the difference between the HighUnexp and HighExp conditions) was reduced when listening to speech in noise. This is in agreement with previous research that has found a decrease in the amplitude of N400 effects when listening to acoustically challenging speech (see Goslin et al., 2012; Obleser & Kotz, 2011; Romero-Rivas et al., 2016; Strauß et al., 2013). Typically, reductions in N400 effects have been viewed as deficiencies in being able to use context to facilitate semantic processing (e.g., Wlotko, Lee, & Federmeier, 2010; Ng et al., 2017, 2018). The reduction in the N400 expectancy effect seen in our data likely suggests that when listening in noise, the listener's ability to use contextual information to build up expectations to facilitate semantic retrieval of individual words is reduced (for a discussion on how the N400 reflects semantic memory retrieval processes see, Kutas & Federmeier, 2000, 2011). Alternatively, under a semantic integration account (Hagoort, Baggio, & Willems, 2009), the reduced N400 could reflect a decreased efficiency in being able to integrate the sentence-final word with the preceding context in noise. Importantly, under either account, this decrease in the expectancy effect clearly reflects that the listener is unable to use contextual information as efficiently in real time to the same degree as when listening in quiet.

We also found that the onset of the N400 expectancy effect was delayed by about 73 msec when listening in noise. Prior work has shown that words with a similar phonological onset to an expected word typically have a delayed onset on the N400 response as compared to words that have an unexpected phonological onset, because the listener had likely built up expectations for phonological features of a predicted word (see Van Petten et al., 1999). During the initial phoneme (the smallest meaningful unit of speech) of the critical word, the listener monitors the acoustic signal for phonological features that match expectations. When these expectations are not met, the system begins a rapid onset of the N400 (Van Petten et al., 1999; Nieuwland, 2019). Others have argued that this type of early onset of a negative deflection to unexpected words represents a neurally distinct ERP component referred to as either a phonological mismatch negativity, phonological mapping negativity, or the auditory N200 component (Boudewyn, Long, & Swaab, 2015; Connolly & Phillips, 1994; Hagoort & Brown, 2000; Van Den Brink, Brown, & Hagoort, 2001). However, regardless of whether or not these early negative deflections represent a distinct component or an earlier onset of the N400 response (for a discussion on this see Nieuwland, 2019), this early ERP effect is delayed (or reduced) if the initial phoneme matches the phonological onset of an expected word. Thus, one could argue that the onset of this N400-like effect in the auditory domain reflects, in part, context-driven predictions of phonological features of upcoming words.

In addition to a reduction in N400 effect amplitude, it has been commonly reported that there is a delay in the latency of either the onset or the peak of the N400 (or a reduction in the phonological mismatch negativity) when listening to perceptually challenging speech (Aydelott et al., 2006; Connolly, Phillips, Stewart, & Brake, 1992; Goslin et al., 2012; Obleser & Kotz, 2011; Strauß et al., 2013). The delays seen in these studies and our data suggests that phonological predictive processes may be impaired when listening to speech in noise. Therefore, the delay we observed may suggest that when listening in noise, instead of using contextual information to predict possible phonological features of upcoming words, the listener may enter into more of a bottom-up "wait and see" mode (Federmeier et al., 2007). This would require the listener to accumulate more phonological information than usual when listening to degraded speech before they begin the processes associated with accessing the semantic information of a particular word (for supporting behavioral evidence see Lash et al., 2013; Nooteboom & Doodeman, 1984).

4.3. *Acoustic challenge and the pupillometric response*

We found that there was a larger pupillary response when listening to speech in noise compared to quiet, consistent with prior work (for a review see Zekveld et al., 2018). The task-evoked pupillary response has previously been used as a marker for locus coeruleus-norepinephrine activity and task-related arousal (e.g., Aston-Jones & Cohen, 2005; Joshi et al., 2016; Reimer et al., 2016; Murphy et al., 2014). It is possible that the increase in pupil size seen for speech heard in noise could reflect changes in arousal and attention, consistent with predictions made by FUEL (Pichora-Fuller et al., 2016).

It should be noted that the SNR used in the current study was relatively higher than past pupillometry studies of listening effort. Much of the work that has been done previously has used individualized SNRs based on individual word intelligibility level (e.g., Zekveld et al., 2010). Because of this, it may be difficult to differentiate between the effects of the masking of the speech and the effects of the induction of *effort* in response to the masking, since these two effects would be highly correlated. However, because participants in our study were able to perceive the speech at near 100% accuracy, the effects we saw on the pupillary response could not be explained directly by the masking of the speech but more likely by the increase in effort experienced by the listeners (see also Kuchinsky et al., 2013; McLaughlin & Van Engen, 2020). Therefore, our results provide strong evidence that the pupillary response is highly sensitive to changes in listening effort, not just changes in intelligibility.

4.4. *The costs and benefits of effortful listening*

Importantly, the results discussed thus far show how the use of context may vary as a function of listening in noise *generally*, but do not tell us how context use may vary as a function of listening *effort* while hearing speech in noise. A key component of listening effort theories (Pichora-Fuller et al., 2016; Peelle, 2018; Zekveld et al., 2018; Rabbitt, 1968, 1991) is that effortful listening is not just about how much perceptual challenge a person is experiencing but also about how a listener *responds* to that challenge. Peelle (2018) recently argued that "in contrast to cognitive demand, listening effort refers to the resources or energy *actually* used by a listener to meet cognitive demand" (p. 205, emphasis added). Indeed, all of our noise trials used a consistent level of perceptual challenge, but it is likely that a listener's motivational, attentional, or arousal state may have varied from trial to trial, leading to varying degrees of effort allocation. Therefore, a major innovation of this study was to use single-trial pupillary changes as a physiological marker of variation in effort allocation (e.g., Zekveld et al., 2018) rather than inferring listening effort from acoustic challenge alone. We used a novel analysis approach that allowed us to examine the relationship between single-trial variability in the pupil size changes and the ERP responses and memory outcomes while listening in noise. This allowed us to use these relationships to differentiate the effects of acoustic challenge from the effects of listening effort.

We found that the amplitude of the N400 expectancy effect (i.e., the difference between the N400 response to expected and unexpected words heard in sentences with highly constraining context) increased as the pupillary response increased, suggesting that increased effort allocation predicts recovery of the use of contextual information to facilitate online semantic processing. This effect can be further illustrated by comparing effect sizes of the N400 expectancy effect under differing conditions. For instance, the N400 overall expectancy effect *in quiet* was approximately $-2.76\ \mu V$. At the average pupillary response *in noise*, this effect was estimated to be $-1.67\ \mu V$, showing a general expectancy reduction in noise. Importantly however, when participants showed larger pupillary responses (1 *SD* above their average response) the model-predicted N400 expectancy effect was $-3.07\ \mu V$. This suggests that the negative effects of noise on the N400 expectancy effect (as has been found previously; e.g., Obleser & Kotz, 2011) can be overcome when participants exert greater listening effort.

In contrast to these findings, we found that larger pupillary responses were associated with poorer performance for *both* sentence recognition memory and word recall, effects that were not modulated by the contextual information available to the listener. Despite the contextual benefits to memory when experiencing acoustic challenge discussed above, the negative main effect of the pupillary response suggests that when listeners expend increased effort when listening to acoustically degraded speech, this effort expenditure had a generally negative effect on memory encoding and later retrieval at both the word and sentence level.

Taken together with the N400 data, these findings are in line with key predictions from listening effort theories, which predict such a tradeoff between online lexical processing and subsequent memory. As a listener exerts more effort to support on-line word recognition processes in the face of degraded speech, fewer resources are available for higher-level memory encoding processes (Rabbitt, 1968, 1991). The innovation of the current study is that by directly and simultaneously measuring physiological markers of effort allocation (i.e., pupil dilation) and real time word processing (i.e., the N400) along with subsequent memory, and directly examining their trial-to-trial covariability, we were able to directly quantify this resource trade off as listening effort changes, giving a more direct window into the mechanisms underlying the concept of resource allocation, which has often been elusive and controversial (e.g., Hommel et al., 2019; Logan, 1988; Navon, 1984).

4.5. *Future directions and conclusions*

In this section, we discuss important caveats of the current study and areas for future work. First, the current study did not include an online behavioral task while participants listened to speech stimuli, obviating our ability to look at correlations with real-time comprehension performance. Importantly however, previous research has shown that different kinds of behavioral tasks can have a direct impact on online measures of language processing, including response time measures of speech processing (e.g., Fallon, Peelle, & Wingfield, 2006, pp. P10–P17) and language-related ERPs (e.g., Payne, Stites, & Federmeier, 2019; Schacht, Sommer, Shmuilovich, Martíenz, & Martín-Loeches, 2014). As such, our goal in the current study was to establish a task-free baseline of the effects on the N400, pupillary responses, and memory outcomes. However, future work should explore the effects that different behavioral tasks might have on these different measures.

Another important caveat is that the context manipulations used in the current study were extreme in nature (i.e., the high context sentences with expected sentence-final words had an average cloze probability of .88 and sentences with unexpected sentence-final words had an average of .01). While this is a canonical manipulation of context used in both the electrophysiology and audiology literatures (e.g., Federmeier et al., 2007; Gordon-Salant & Fitzgibbons, 1997), there is a growing body of research that has shown that the N400, in particular, shows a graded response to sentential constraint (e.g., Payne & Federmeier, 2017b; Wlotko & Federmeier, 2012). Therefore, future work should assess how continuous manipulations of constraint or cloze probability may influence the effects that we saw.

In conclusion, the Framework for Understanding Effortful Listening proposed by Pichora-Fuller et al. (2016) posits that the allocation of limited resources depends upon *both* the cognitive demand of the task at hand *and* the way that the listener responds to that demand (see also Peelle, 2018; Zekveld et al., 2018). The findings from the current study clearly demonstrated direct evidence for this claim, as noise-induced acoustic challenge effects (i.e., cognitive/neural demand) and pupil-mediated effort effects showed distinct and divergent influences on both real-time neural measures of word processing and subsequent memory performance.

Author contributions

Jack W Silcox: Conceptualization, Data Curation, Formal Analysis, Writing-original draft, Writing-review & editing.

Brennan R Payne: Conceptualization, Project administration, Writing-review & editing.

Author note

Portions of this work were presented at the Society for Psychophysiological Research (2020). We thank the Office of the Vice President for Research and the Office of Undergraduate Research at the University of Utah for providing research support for this project. We thank Jordan Anderson, Elisabeth Antley, Karen Bennett, Hannah Crandell, Brett Smith, Jessica Stoker, and Mary Yoo for assistance in data collection.

Open practices

The study in this article earned Open Data and Preregistered badges for transparent practices. Data for this study is available at: https://osf.io/hcrv6/files/.

REFERENCES

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage Publications.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28*, 403–450.
Aydelott, J., Dick, F., & Mills, D. L. (2006). Effects of acoustic distortion and semantic context on event-related potentials to spoken words. *Psychophysiology, 43*(5), 454–464.
Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., et al. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.
Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). *Parsimonious mixed models*. arXiv preprint arXiv:1506.04967.
Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4*. arXiv preprint arXiv:1406.5823.
Benichov, J., Cox, L. C., Tun, P. A., & Wingfield, A. (2012). Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability and cognitive function. *Ear and Hearing, 32*(2), 250.
Benton, A. L., Hamsher, K. D., & Sivan, A. B. (1978). *Multilingual aphasia examination: Manual of instructions*. IA City, IA: AJA Assoc.
Berrien, F. K., & Huntington, G. H. (1943). An exploratory study of pupillary responses during deception. *Journal of Experimental Psychology, 32*(5), 443.
Blackburn, K., & Schirillo, J. (2012). Emotive hemispheric differences measured in real-life portraits using pupil diameter and subjective aesthetic preferences. *Experimental Brain Research, 219*(4), 447–455.
Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods, 42*(3), 665–670.
Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience, 15*(3), 607–624.
Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology, 45*(4), 602–607.
Brehm, J. W., & Self, E. A. (1989). The intensity of motivation. *Annual Review of Psychology, 40*(1), 109–131.
Brehm, J. W., Wright, R. A., Solomon, S., Silka, L., & Greenberg, J. (1983). Perceived difficulty, energization, and the magnitude of goal valence. *Journal of Experimental Social Psychology, 19*(1), 21–48.
Breton-Provencher, V., & Sur, M. (2019). Active control of arousal by a locus coeruleus GABAergic circuit. *Nature Neuroscience, 22*(2), 218–228.
Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990.
Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911.
Connolly, J. F., & Phillips, N. A. (1994). Event-related potential components reflect phonological and semantic processing of the terminal word of spoken sentences. *Journal of Cognitive Neuroscience, 6*(3), 256–266.
Connolly, J. F., Phillips, N. A., Stewart, S. H., & Brake, W. G. (1992). Event-related potential sensitivity to acoustic and semantic properties of terminal words in sentences. *Brain and Language, 43*(1), 1–18.
Cousins, K. A., Dar, H., Wingfield, A., & Miller, P. (2014). Acoustic masking disrupts time-dependent mechanisms of memory encoding in word-list recall. *Memory & Cognition, 42*(4), 622–638.
Craik, F. I. (2002). Levels of processing: Past, present... and future? *Memory, 10*(5–6), 305–318.
Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684.
Craik, F. I., & McDowd, J. M. (1987). Age differences in recall and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13*(3), 474.
Danckert, S. L., & Craik, F. I. (2013). Does aging affect recall more than recognition memory? *Psychology and Aging, 28*(4), 902.
Ekstrom, R. B., Dermen, D., & Harman, H. H. (1976). *Manual for kit of factor-referenced cognitive tests* (Vol. 102). Princeton, NJ: Educational testing service.
Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12*(2), 121.
Fallon, M., Peelle, J. E., & Wingfield, A. (2006). Spoken sentence processing in young and older adults modulated by task demands: Evidence from self-paced listening. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 61*(1), P10–P17.
Federmeier, K. D., Kutas, M., & Schul, R. (2010). Age-related and individual differences in the use of prediction during language comprehension. *Brain and Language, 115*(3), 149–161.
Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain Research, 1146*, 75–84.
Ferreira, F., Bailey, K. G., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science, 11*(1), 11–15.
Ferreira, F., & Lowder, M. W. (2016). Prediction, information structure, and good-enough language processing. In *Psychology of learning and motivation* (Vol. 65, pp. 217–247). Academic Press.
Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass, 1*(1-2), 71–83.
Gordon-Salant, S., & Fitzgibbons, P. J. (1997). Selected cognitive factors and speech recognition performance among young and elderly listeners. *Journal of Speech, Language, and Hearing Research, 40*(2), 423–431.
Goslin, J., Duffy, H., & Floccia, C. (2012). An ERP investigation of regional and foreign accent processing. *Brain and Language, 122*(2), 92–102.
Guang, C., Lefkowitz, E., Dillman-Hasso, N., Brown, V., & Strand, J. (2021). Recall of speech is impaired by subsequent masking noise: A replication of Rabbitt (1968) experiment 2. *Auditory Perception & Cognition*, 1–10.
Hagoort, P., Baggio, G., & Willems, R. M. (2009). Semantic unification. In *The cognitive neurosciences* (4th ed., pp. 819–836). MIT press.
Hagoort, P., & Brown, C. M. (2000). ERP effects of listening to speech: Semantic ERP effects. *Neuropsychologia, 38*(11), 1518–1530.
Hess, E. H., & Polt, J. M. (1960). Pupil size as related to interest value of visual stimuli. *Science, 132*(3423), 349–350.
Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190–1192.
Hommel, B., Chapman, C. S., Cisek, P., Neyedli, H. F., Song, J. H., & Welsh, T. N. (2019). No one knows what attention is. Attention, Perception. *& Psychophysics, 81*(7), 2288–2303.
Jacoby, L. L., Craik, F. I., & Begg, I. (1979). Effects of decision difficulty on recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 18*(5), 585–600.
Jacoby, L. L., Toth, J. P., & Yonelinas, A. P. (1993). Separating conscious and unconscious influences of memory: Measuring recollection. *Journal of Experimental Psychology: General, 122*(2), 139.

Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalogr. Clin. Neurophysiol., 10*, 370–375.
Joshi, S., & Gold, J. I. (2020). Pupil size as a window on neural substrates of cognition. *Trends in Cognitive Sciences, 24*(6), 466–480.
Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron, 89*(1), 221–234.
Kiesel, A., Miller, J., Jolicœur, P., & Brisson, B. (2008). Measurement of ERP latency differences: A comparison of single-participant and jackknife-based scoring methods. *Psychophysiology, 45*(2), 250–274.
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science, 3*(2), 346–370.
Knapen, T., de Gee, J. W., Brascamp, J., Nuiten, S., Hoppenbrouwers, S., & Theeuwes, J. (2016). Cognitive and ocular factors jointly determine pupil responses under equiluminance. *Plos One, 11*(5), Article e0155574.
Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research, 323*, 81–90.
Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hearing Research, 312*, 114–120.
Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012a). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291–300.
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. *International Journal of Otolaryngology*, 2012.
Koeritzer, M. A., Rogers, C. S., Van Engen, K. J., & Peelle, J. E. (2018). The impact of age, background noise, semantic ambiguity, and hearing loss on recognition memory for spoken sentences. *Journal of Speech, Language, and Hearing Research, 61*(3), 740–751.
Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods, 51*(3), 1336–1342.
Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Jr., Cute, S. L., Humes, L. E., Dubno, J. R., et al. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*(1), 23–34.
Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences, 4*(12), 463–470.
Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event-related brain potential (ERP). *Annual Review of Psychology, 62*, 621–647.
Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science, 207*(4427), 203–205.
Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*(5947), 161–163.
Lash, A., Rogers, C. S., Zoller, A., & Wingfield, A. (2013). Expectation and entropy in spoken word recognition: Effects of age and hearing acuity. *Experimental Aging Research, 39*(3), 235–253.
Lau, E. F., Phillips, C., & Poeppel, D. (2008). A cortical network for semantics:(de) constructing the N400. *Nature Reviews Neuroscience, 9*(12), 920–933.
Lenth, R., Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). *emmeans: Estimated marginal means, aka least-squares means* Version 1.3. 4.
Logan, G. D. (1988). Automaticity, resources, and memory: Theoretical controversies and practical implications. *Human Factors, 30*(5), 583–598.
Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87*(3), 252.
Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*(5417), 522–523.
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. *The Quarterly Journal of Experimental Psychology Section A, 58*(1), 22–33.
McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology, 54*(2), 193–203.
McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal of the Acoustical Society of America, 147*(2), EL151–EL156.
McMahon, C. M., Boisvert, I., de Lissa, P., Granger, L., Ibrahim, R., Lo, C. Y., et al. (2016). Monitoring alpha oscillations and pupil dilation across a performance-intensity function. *Frontiers in Psychology, 7*, 745.
Murphy, P. R., O'connell, R. G., O'sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human Brain Mapping, 35*(8), 4140–4154.
Navon, D. (1984). Resources—a theoretical soup stone? *Psychological Review, 91*(2), 216–234.
Ng, S., Payne, B. R., Steen, A. A., Stine-Morrow, E. A., & Federmeier, K. D. (2017). Use of contextual information and prediction by struggling adult readers: Evidence from reading times and event-related potentials. *Scientific Studies of Reading, 21*(5), 359–375.
Ng, S., Payne, B. R., Stine-Morrow, E. A., & Federmeier, K. D. (2018). How struggling adult readers use contextual information when comprehending speech: Evidence from event-related potentials. *International Journal of Psychophysiology, 125*, 1–9.
Nieuwland, M. S. (2019). Do 'early'brain responses reveal word form prediction during language comprehension? A critical review. *Neuroscience and Biobehavioral Reviews, 96*, 367–400.
Nooteboom, S. G., & Doodeman, G. J. N. (1984). Speech quality and the gating paradigm. In *Proceedings of the tenth international congress of phonetic sciences* (pp. 481–485) (Foris Dordrecht).
Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600–2606.
Obleser, J., & Kotz, S. A. (2011). Multiple brain signatures of integration in the comprehension of degraded speech. *Neuroimage, 55*(2), 713–723.
Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., et al. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research, 351*, 68–79.
Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*(1), 97–113.
Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods, 47*(4), 1343–1355.
Payne, B. R., & Federmeier, K. D. (2017a). Pace yourself: Intraindividual variability in context use revealed by self-paced event-related brain potentials. *Journal of Cognitive Neuroscience, 29*(5), 837–854.

Payne, B. R., & Federmeier, K. D. (2017b). Event-related brain potentials reveal age-related changes in parafoveal-foveal integration during sentence processing. *Neuropsychologia, 106*, 358–370.

Payne, B. R., & Federmeier, K. D. (2019). Individual differences in reading speed are linked to variability in the processing of lexical and contextual information: Evidence from single-trial event-related brain potentials. *Word, 65*(4), 252–272.

Payne, B. R., Lee, C. L., & Federmeier, K. D. (2015). Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. *Psychophysiology, 52*(11), 1456–1469.

Payne, B. R., & Silcox, J. W. (2019). Aging, context processing, and comprehension. In *Psychology of learning and motivation* (Vol. 71, pp. 215–264). Academic Press.

Payne, B. R., Silcox, J., Crandell, H., Lash, A., Ferguson, S. H., & Lohani, M. (2021). Text captioning buffers against the effects of background noise and hearing loss on memory for speech. *Ear and Hearing*. https://doi.org/10.31234/osf.io/59cq8. in press.

Payne, B. R., Stites, M. C., & Federmeier, K. D. (2019). Event-related brain potentials reveal how multiple aspects of semantic processing unfold across parafoveal and foveal vision during sentence reading. *Psychophysiology, 56*(10), Article e13432.

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.

Pichora-Fuller, M. K. (2008). Use of supportive context by younger and older adult listeners: Balancing bottom-up and top-down information processing. *International Journal of Audiology, 47*(sup2), S72–S82.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., et al. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing, 37*, 5S–27S.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America, 97*(1), 593–608.

Piquado, T., Benichov, J. I., Brownell, H., & Wingfield, A. (2012). The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening. *International Journal of Audiology, 51*(8), 576–583.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology, 20*(3), 241–248.

Rabbitt, P. (1991). Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta oto-laryngologica, 111*(sup476), 167–176.

Reimer, J., McGinley, M. J., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D. A., et al. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications, 7*(1), 1–7.

Richter, J., Gendolla, G. H., & Wright, R. A. (2016). Three decades of research on motivational intensity theory: What we have learned about effort and what we still don't know. *Advances in Motivation Science, 3*, 149–186.

Rogers, C. S. (2017). Semantic priming, not repetition priming, is to blame for false hearing. *Psychonomic Bulletin & Review, 24*(4), 1194–1204.

Rogers, C. S., Jacoby, L. L., & Sommers, M. S. (2012). Frequent false hearing by older adults: The role of age differences in metacognition. *Psychology and Aging, 27*(1), 33.

Romero-Rivas, C., Martin, C. D., & Costa, A. (2016). Foreign-accented speech modulates linguistic anticipatory processes. *Neuropsychologia, 85*, 245–255.

Rommers, J., & Federmeier, K. D. (2018). Predictability's aftermath: Downstream consequences of word predictability as revealed by repetition effects. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 101*, 16–30.

Rugg, M. D., & Yonelinas, A. P. (2003). Human recognition memory: A cognitive neuroscience perspective. *Trends in Cognitive Sciences, 7*(7), 313–319.

Schacht, A., Sommer, W., Shmuilovich, O., Martíenz, P. C., & Martín-Loeches, M. (2014). Differential task effects on N400 and P600 elicited by semantic and syntactic violations. *Plos One, 9*(3), Article e91226.

Schiller, N. O., Boutonnet, B. P. A., De Heer Kloots, M. L., Meelen, M., Ruijgrok, B., & Cheng, L. L. S. (2020). (Not so) Great expectations: Listening to foreign-accented speech reduces the brain's anticipatory processes. *Frontiers in Psychology, 11*, 2143.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303–304.

Sheldon, S., Pichora-Fuller, M. K., & Schneider, B. A. (2008). Priming and sentence context support listening to noise-vocoded speech by younger and older adults. *The Journal of the Acoustical Society of America, 123*(1), 489–499.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2015). *afex: Analysis of factorial experiments. R package version 0.13–145*.

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(6), 679–692.

Smeds, K., Wolters, F., & Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *Journal of the American Academy of Audiology, 26*(2), 183–196.

Strauß, A., Kotz, S. A., & Obleser, J. (2013). Narrowed expectancies under degraded speech: Revisiting the N400. *Journal of Cognitive Neuroscience, 25*(8), 1383–1395.

Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of Clinical Neuropsychology, 14*(2), 167–177.

Tun, P. A., O'Kane, G., & Wingfield, A. (2002). Distraction by competing speech in young and older adult listeners. *Psychology and Aging, 17*(3), 453.

Ulrich, R., & Miller, J. (2001). Using the jackknife-based scoring method for measuring LRP onset effects in factorial designs. *Psychophysiology, 38*(5), 816–827.

Van Den Brink, D., Brown, C. M., & Hagoort, P. (2001). Electrophysiological evidence for early contextual influences during spoken-word recognition: N200 versus N400 effects. *Journal of Cognitive Neuroscience, 13*(7), 967–985.

Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *Plos One, 7*(9), Article e43753.

Van Gerven, P. W., Paas, F., Van Merriënboer, J. J., & Schmidt, H. G. (2004). Memory load and the cognitive pupillary response in aging. *Psychophysiology, 41*(2), 167–174.

Van Petten, C., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*(2), 394.

Van Petten, C., & Luka, B. J. (2006). Neural localization of semantic context effects in electromagnetic and hemodynamic studies. *Brain and Language, 97*(3), 279–293.

Varazzani, C., San-Galli, A., Gilardeau, S., & Bouret, S. (2015). Noradrenaline and dopamine neurons in the reward/effort trade-off: A direct electrophysiological comparison in behaving monkeys. *Journal of Neuroscience, 35*(20), 7866–7877.

Wagner, A. E., Toffanin, P., & Başkent, D. (2016). The timing and effort of lexical access in natural and degraded speech. *Frontiers in Psychology, 7*, 398.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods, 45*(4), 1191–1207.

Webb, A. K., Hacker, D. J., Osher, D., Cook, A. E., Woltz, D. J., Kristjansson, S., et al. (2009, July). Eye movements and pupil size reveal deception in computer administered questionnaires. In *International conference on foundations of augmented cognition* (pp. 553–562). Berlin, Heidelberg: Springer.

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

Westfall, J. (2015). *PANGEA: Power analysis for general ANOVA designs*. Unpublished manuscript. Available at http://jakewestfall.org/publications/pangea.pdf.

Wilson, M. (1988). MRC psycholinguistic database: Machine-useable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers, 20*(1), 6–10.

Wingfield, A., Tun, P. A., & McCoy, S. L. (2005). Hearing loss in older adulthood: What it is and how it interacts with cognitive performance. *Current Directions in Psychological Science, 14*(3), 144–148.

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*, 2331216516669723.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing, 36*(4), e153.

Winneke, A. H., Schulte, M., Vormann, M., & Latzel, M. (2020). Effect of directional microphone technology in hearing aids on neural correlates of listening and memory effort: An electroencephalographic study. *Trends in Hearing, 24*, 2331216520948410.

Winn, M. B., & Teece, K. H. (2021). *Listening effort is not the same as speech intelligibility score*. https://doi.org/10.31234/osf.io/vk65w. PsyArXiv [Preprint]. Available from.

Wlotko, E. W., & Federmeier, K. D. (2007). Finding the right word: Hemispheric asymmetries in the use of sentence context information. *Neuropsychologia, 45*(13), 3001–3014.

Wlotko, E. W., & Federmeier, K. D. (2012). So that's what you meant! Event-related potentials reveal multiple aspects of context use during construction of message-level meaning. *Neuroimage, 62*(1), 356–366.

Wlotko, E. W., Lee, C. L., & Federmeier, K. D. (2010). Language of the aging brain: Event-related potential studies of comprehension in older adults. *Language and Linguistics Compass, 4*(8), 623–638.

Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., & Oleson, J. (2018). Characteristics of real-world signal-to-noise ratios and speech listening situations of older adults with mild-to-moderate hearing loss. *Ear and Hearing, 39*(2), 293.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517.

Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2013). Task difficulty differentially affects two measures of processing load: The pupil response during sentence processing and delayed cued recall of the sentences. *Journal of Speech, Language, and Hearing Research, 56*, 1156–1165.

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014a). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *Neuroimage, 101*, 76–86.

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

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology, 51*(3), 277–284.

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing, 31*(4), 480–490.

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 and Hearing, 32*(4), 498–510.

Zekveld, A. A., Rudner, M., Kramer, S. E., Lyzenga, J., & Rönnberg, J. (2014b). 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.