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

ABSTRACT

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.

© 2021 Elsevier Ltd. All rights reserved.
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. 105; 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 audiological 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 texts, particularly for adults with hearing impairment. was presented. They found that all participants showed 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 1980s (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, 2008). 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, & Flocia, 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; Siros & 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; Siros & Brisson, 2014; Van Gerven, Paas, Van Merrienboer, & Schmidt, 2004), motivation (Knappen 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, Księżek, 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” (Fichora-Fuller, 2008, p. 575, 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, Strauß 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, Wlooto, 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 (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 righthanded 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. 

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 (Low-Unexp): 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 (SUBTLEXus 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, Rönning, & Kramer, 2012; 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 (Smits, 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 and 900 msec after 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 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 sentence audio. This time window also 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. 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, Burker, & 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, Joliceur, & 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.
2.8. 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. 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.

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 to 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 ($\chi^2 (2) = 41.43, p < .01$) but no main effect of noise ($\chi^2 (1) = .19, p = .67$). However, these effects were qualified by a significant interaction between word type and noise ($\chi^2 (2) = 6.09, p < .05$). 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 High-Unexp that was marginally significant. Importantly, we found that while the mean N400 amplitude was larger for High-Unexp than for HighExp both in quiet and in noise, this effect was significantly reduced in magnitude in noise compared to quiet.
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 ($t_{cor}$ = 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. 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.
constraint sentences, recognition was significantly worse in 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 ($\chi^2 (2) = 213.07, p < .01$). Contrasts showed that HighExp words were remembered better than both HighUnexp and LowUnexp words. Moreover, there was a main effect of background noise ($\chi^2 (1) = 5.60, p < .05$), 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 ($\chi^2 (2) = .52, p = .77$).

### 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 ($\chi^2 (1) = 9.59, p < .01$), 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 ($\chi^2 (2) = 33.47, p < .01$) but no main effect of pupil size ($\chi^2 (1) = .30, p = .58$) on single trial N400 amplitude. However, there was a significant interaction between target word type and pupil size ($\chi^2 (2) = 12.60, p < .01$). 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.

---

### 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.

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Diff. Est. t(df)</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quiet: HighExp versus HighUnexp</td>
<td>2.76</td>
<td>.03</td>
<td>[1.46, 4.06]</td>
</tr>
<tr>
<td>Noise: HighExp versus HighUnexp</td>
<td>1.64</td>
<td>&lt;.01</td>
<td>[26, 3.01]</td>
</tr>
<tr>
<td>Quiet: HighExp versus LowUnexp</td>
<td>1.75</td>
<td>&lt;.01</td>
<td>[29, 3.21]</td>
</tr>
<tr>
<td>Noise: HighExp versus LowUnexp</td>
<td>2.07</td>
<td>&lt;.01</td>
<td>[74, 3.41]</td>
</tr>
<tr>
<td>Quiet: HighUnexp versus LowUnexp</td>
<td>-1.01</td>
<td>.15</td>
<td>[-2.81, .79]</td>
</tr>
<tr>
<td>Noise: HighUnexp versus LowUnexp</td>
<td>.44</td>
<td>.39</td>
<td>[-92, 1.80]</td>
</tr>
<tr>
<td>HighExp: Quiet versus Noise</td>
<td>-.16</td>
<td>.90</td>
<td>[-1.49, 1.18]</td>
</tr>
<tr>
<td>HighUnexp: Quiet versus Noise</td>
<td>.97</td>
<td>.05</td>
<td>[-34, 2.28]</td>
</tr>
<tr>
<td>LowUnexp: Quiet versus Noise</td>
<td>.48</td>
<td>.38</td>
<td>[-1.79, .83]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Diff. Est. t(df)</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Constraint: Quiet versus Noise</td>
<td>-.01</td>
<td>.84</td>
<td>[-.07, .06]</td>
</tr>
<tr>
<td>Low Constraint: Quiet versus Noise</td>
<td>.08</td>
<td>&lt;.05</td>
<td>[.02, .15]</td>
</tr>
<tr>
<td>Quiet: High versus Low Constraint</td>
<td>.22</td>
<td>&lt;.01</td>
<td>[.16, .29]</td>
</tr>
<tr>
<td>Noise: High versus Low Constraint</td>
<td>.31</td>
<td>&lt;.01</td>
<td>[.25, .37]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Contrast</th>
<th>Diff. Est. t(df)</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighExp versus HighUnexp</td>
<td>.34</td>
<td>&lt;.01</td>
<td>[28, 40]</td>
</tr>
<tr>
<td>HighExp versus LowUnexp</td>
<td>.48</td>
<td>&lt;.01</td>
<td>[42, 54]</td>
</tr>
<tr>
<td>HighUnexp versus LowUnexp</td>
<td>.14</td>
<td>&lt;.01</td>
<td>[08, 20]</td>
</tr>
<tr>
<td>Quiet versus Noise</td>
<td>.05</td>
<td>&lt;.05</td>
<td>[.01,0.09]</td>
</tr>
</tbody>
</table>
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 ($t_{\text{corrected}} (38) = -0.46$, $p = 0.65$). 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 0.74 (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 ($\chi^2 (1) = 40.08$, $p < .01$) 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 ($\chi^2 (1) = 4.10$, $p < .05$) 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 ($\chi^2 (1) = 2.01$, $p = .16$).

The results from the recall portion of the memory test were similar. We found that there was a main effect of target word type ($\chi^2 (2) = 196.30$, $p < .01$) with the odds of recalling an
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 ($\chi^2 (1) = 5.68, p < .05$), 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 ($\chi^2 (2) = 2.53, p = .28$).

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
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.

**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>
<thead>
<tr>
<th>Simple Slope Estimates</th>
<th>Estimate</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Constraint, Expected</td>
<td>.69</td>
<td>[.13, 1.24]</td>
</tr>
<tr>
<td>High Constraint, Unexpected</td>
<td>-1.27</td>
<td>[-1.27, -1.15]</td>
</tr>
<tr>
<td>Low Constraint, Unexpected</td>
<td>.30</td>
<td>[-.28, .88]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Simple Slope Contrasts</th>
<th>Diff. Est.</th>
<th>z</th>
<th>p-value</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>HighExp versus HighUnexp</td>
<td>1.40</td>
<td>3.45</td>
<td>&lt;.01</td>
<td>[.43, 2.36]</td>
</tr>
<tr>
<td>HighExp versus LowUnexp</td>
<td>.39</td>
<td>.95</td>
<td>.34</td>
<td>[-.60, 1.37]</td>
</tr>
<tr>
<td>HighUnexp versus LowUnexp</td>
<td>-1.01</td>
<td>-2.45</td>
<td>&lt;.05</td>
<td>[-1.99, -.02]</td>
</tr>
</tbody>
</table>
trade-off between word processing and subsequent memory, consistent with the predictions of listening effort theories, such as FUEL (e.g., Rabbitt, 1968, 1991; Fichora-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 et al., 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, & Younélis, 1993, 1979; Mandler, 1980; Rugg & Younélis, 2003; Younélis, 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 Kutu & 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, Steine-Morrow, & Federmeier, 2017; Payne & Federmeier, 2017a, 2019; Włotko & 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; Olesler & 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., Włotko, Lee, & Federmeier, 2016; 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; Nootbooom & 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; Rabett, 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 μV. At the average pupillary response in noise, this effect was estimated to be −1.67 μV, 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 μV. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


duct such a tradeoff between online lexical processing and subsequent memory performance. 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 (Rabbit, 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 win-


