OPEN

The Effort of Repairing a Misperceived Word Can Impair Perception of Following Words, Especially for Listeners With Cochlear Implants

Matthew B. Winn[1]

Objectives: In clinical and laboratory settings, speech recognition is typically assessed in a way that cannot distinguish accurate auditory perception from misperception that was mentally repaired or inferred from context. Previous work showed that the process of repairing misperceptions elicits greater listening effort, and that this elevated effort lingers well after the sentence is heard. That result suggests that cognitive repair strategies might appear successful when testing a single utterance but fail for everyday continuous conversational speech. The present study tested the hypothesis that the effort of repairing misperceptions has the consequence of carrying over to interfere with perception of later words after the sentence.

Design: Stimuli were open-set coherent sentences that were presented intact or with a word early in the sentence replaced with noise, forcing the listener to use later context to mentally repair the missing word. Sentences were immediately followed by digit triplets, which served to probe carryover effort from the sentence. Control conditions allowed for the comparison to intact sentences that did not demand mental repair, as well as to listening conditions that removed the need to attend to the post-sentence stimuli, or removed the post-sentence digits altogether. Intelligibility scores for the sentences and digits were accompanied by time-series measurements of pupil dilation to assess cognitive load during the task, as well as subjective rating of effort. Participants included adults with cochlear implants (CIs), as well as an age-matched group and a younger group of listeners with typical hearing for comparison.

Results: For the CI group, needing to repair a missing word during a sentence resulted in more errors on the digits after the sentence, especially when the repair process did not result in a coherent sensible perception. Sentences that needed repair also contained more errors on the words that were unmasked. All groups showed substantial increase of pupil dilation when sentences required repair, even when the repair was successful. Younger typical hearing listeners showed clear differences in moment-to-moment allocation of effort in the different conditions, while the other groups did not.

Conclusions: For CI listeners, the effort of needing to repair misperceptions in a sentence can last long enough to interfere with words that follow the sentence. This pattern could pose a serious problem for regular communication but would go overlooked in typical testing with single utterances, where a listener has a chance to repair misperceptions before responding. Carryover effort was not predictable by basic intelligibility scores, but can be revealed in behavioral data when sentences are followed immediately by extra probe words such as digits.

[1]Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, Minnesota, USA.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and text of this article on the journal's Web site (www.ear-hearing.com).

Key words: Cochlear implants, Listening effort, Pupillometry, Speech perception.

(Ear & Hearing 2024;45;1527–1541)

This article has received OSF badges for Open Data and Open Material.

INTRODUCTION

A major methodological limitation in audiology and speech perception research is that most outcome measures are the end-product of an entire chain of cognitive processing, where accurate auditory perception cannot be distinguished from misperception that was mentally repaired or inferred from context. The common practice of quantifying success by measuring repetition accuracy for sentences neglects the increased effort involved in the mental repair process, and also neglects the downstream implications for the listener's readiness to hear the next sentence. Considering the greater likelihood of auditory perceptual errors by people who are deaf or hard-of-hearing, there could be continual risk of carryover effort from one sentence to the next. The present study is designed to address this important issue by implementing a testing design where listeners must mentally repair misperceptions while continuing to listen to ongoing speech immediately after the sentence. Following recent work on this topic, the present study specifically examines individuals who use cochlear implants (CIs), assuming that the testing method will apply more broadly to others who are deaf or hard-of-hearing.

CIs are famously successful for improving speech communication, but CI users still have hearing difficulties worthy of recognition and further understanding. The effort of listening with a CI can lead to severe psychological barriers to communication and social participation, as reported directly from patients with CIs (Hughes et al. 2018). The social aspects of hearing are linked strongly with effort (McRackan et al. 2017; Hughes et al. 2021) and also predict negative psychological symptoms better than audiometric factors (Crowson et al. 2021). Elevated listening effort has been linked to increased fatigue, mental strain, burnout, medical sick leave, and increased time to recover from daily activities (Kramer et al. 2006; Nachtegaal et al. 2009, 2012), underscoring how listening effort extends far beyond audition to affect quality of life in important ways.

The effort of mentally repairing misperceptions is likely to be underappreciated if outcome measures focus mainly on a tally on the percentage of words correctly repeated. Unsurprisingly, the importance of mental repair has been recognized in popular frameworks of listening effort such as the Ease of Language Understanding model (Rönnberg et al. 2019), which highlights

<zdoi; 10.1097/AUD.0000000000001537>

the concept of "postdiction" as a key component of cognitive load when listening to speech. Winn and Teece (2022) validated that notion in an experiment that presented listeners with sentence stimuli where key target words were completely masked by noise, but recoverable from later context (distinct from perceptual restoration, where the listener has an illusory experience of having actually perceived the original word or phoneme that was masked—see Warren 1970; Balling et al. 2017). In the study by Winn and Teece, the "correct" response could only have occurred because of mental repair because words were not predictable from previous context. This repair process resulted in increased effort even if the full sentence was repeated correctly, for both listeners with typical hearing (TH, defined as pure-tone thresholds better than 25 dB HL) (Winn & Teece 2021) and listeners with CIs (Winn & Teece 2022). Notably, the effort of repairing the missing target word was often maladaptive for the CI listeners; even when the missing word was repaired, there was a greater likelihood to make mistakes elsewhere in the sentence, suggesting a spreading burden of effort that partly motivates the present work.

In the present study, we focus on the possibility that mentally repairing misperceptions could invoke effortful cognitive processes that persist for long enough to impair perception of the next utterance that would be heard. The carryover impact from one sentence to the next is not recognized in standard clinical and laboratory tests of speech recognition, which typically use a single utterance per trial, allowing the listener to freely engage any mental repair processes without any competing demands. Although this testing method enables easier quantification of intelligibility and can avoid complications of remembering long stimuli, results can give the appearance of successful hearing because of a strategy that would not be feasible in real-time conversation. Szostak and Pitt (2013) show that words impact perception of later words downstream, suggesting that perception tasks should account for long-lasting processing time.

Numerous studies have highlighted the importance of extra moments of processing time that support speech comprehension. (Wingfield et al. 1999) found that performance deficits for older compared with younger listeners were neutralized with the introduction of extra silence at syntactic boundaries during the speech, suggesting that the age-related deficits resulted from lack of sufficient processing time. Listeners will adjust incoming speech to add pauses when given the opportunity (Piquado et al. 2012), and this alteration can partially reduce the effort of listening to longer passages (O'Leary et al. 2023). Natural pauses during continuous speech also allow a listener to handle more severe amounts of time compression, and this effect is most clearly observed in listeners who have hearing loss (Schurman, Reference Note 1). Even single sentences are recognized more successfully when followed by moments of silence versus moments of noise disruption—even when the sentence itself is free from any noise masking (Gianakas et al. 2022). Collectively these studies support the notion that moments of processing time during speech comprehension can play a vital role in listener success.

Carryover effects from one utterance to the next have rarely been examined directly. In one recent study, Svirsky et al. (2024) presented listeners with two sentences sequentially instead of just one, finding that the detrimental impact on two-sentence repetition performance ranged up to 45% points. In that study, more than half of CI listeners' scores dropped by 10 or more percentage points, and over a third declined by more than 20 points when adding the second sentence. Notably, the impact of the second sentence was not predictable from the CI users' performance in standard clinical tests like word recognition and AzBio sentence scores, suggesting that carryover effort should be measured directly rather than being inferred from some other outcome measure.

Design of a Task to Measure Carryover Effort

The present study introduces an outcome metric for carryover effort that avoids the potential memory/semantic complications of sequencing two sentences, and avoids the need for discourse-length stimuli that might constrain the granularity of a performance measure (i.e., a small number of comprehension questions). Carryover effort is expected to not merely lower performance overall, but to specifically jeopardize perception at the moment immediately after an effortful event. This approach was previously attempted by Winn and Moore (2018), who introduced a random digit triplet after high- or low-context sentences. They found poorer performance in digit recognition following low-context sentences, which was consistent with prolonged elevated pupil dilations elicited by those sentences (Winn 2016). However, that study could not provide definitive proof that any mental repair took place during the sentence, as a repaired stimulus could not be differentiated from one that was simply heard correctly. Combining methods from multiple aforementioned studies, the present experiment uses sentences that were designed prospectively to demand mental repair, followed immediately by a digit triplet to probe whether the effort of that report carried forward in time.

Although digit sequences are unlikely to be a good probe for listening effort on their own (because they are perceived so reliably well compared with other auditory stimuli), they are an attractive probe stimulus for carryover effort, because they are unlikely to semantically interact with the sentence, and because digit errors should reflect cognitive interference rather than auditory misperception. Digits can be placed at a specific time because they are relatively short in duration, and therefore can be used to examine cognitive load at a specific moment during a listening task.

Previous work has used digits before a sentence as an extra layer of cognitive load, as measured by response times for the sentence (Hunter 2021) as well as accuracy for the digits serving as the pre-sentence load (Hunter & Pisoni 2018). In those studies, pre-sentence cognitive load interfered with the benefit of contextual semantic cues in the sentence. In the present study, we expand on that work by exploring whether extra load of processing a sentence—specifically when it invokes context-driven mental repair—will impair the processing of extra digit stimuli presented immediately after the sentence. Toward understanding the potential effect of the mere presence of the digits as an additional burden, the present study includes a variety of conditions, including one where digits are to be repeated along with the sentence, a condition where the digits are to be ignored, and a condition where there are no digits—just silence after the sentence.

Hypotheses

There were several hypotheses for the present study based on the previous work.

1. When a sentence demands mental repair, we expected listeners to make more errors on the digit triplet that immediately follows that sentence. If true, this would suggest

that the process of mental repair will carry over to compete with ongoing auditory attention, but if false, it would suggest a protected memory buffer that accommodates mental repair without compromising ongoing listening.

2. We expected to replicate findings relating to intelligibility established by Winn and Teece (2022)—specifically that sentences demanding mental repair would yield more errors on words other than the repaired target word, and would yield more responses that are linguistically incoherent. This would suggest that mental repair and ongoing listening/language processing demand shared cognitive resources, that is, the effort devoted to repair would otherwise be used to attend to the rest of the sentence and generate a feasible construction from an incomplete perception. Failure to observe this result would suggest that the words in the sentences are perceived essentially independently.
3. We expected that perception would be more effortful (will elicit greater pupil dilation) when a sentence demands mental repair, even when the listener successfully figures out the missing word. This would suggest that the effort of sentence perception would be governed more strongly by the process of generating feasible candidate perceptions, whereas failure to observe this result would suggest that effort is governed more strongly by ultimate accuracy in repetition.
4. Because listeners are expected to preserve some cognitive resources to hear upcoming stimuli, sentence repetition accuracy was expected to decrease when listeners were required to attend to digits after the sentence. If true, this result would suggest that anticipatory attention can disrupt ongoing listening, but if this result were not found, it would suggest that the sentence is given priority over the digits.

MATERIALS AND METHODS

Participants

The study included three groups of listeners, including 21 young adults with TH thresholds, 22 adults with CIs, and 18 adults with TH who were approximately age-matched to the CI group (see Fig. 1 for a display of participant ages). Audiometric criteria for the TH group were bilateral pure-tone thresholds at or better than 25 dB HL at octave frequencies between 250 and 8000 Hz, with thresholds of 35 dB at 8000 Hz accepted for 2 older listeners. Two younger TH adults and 1 CI listener were excluded because of data quality (see Pupil data preprocessing).

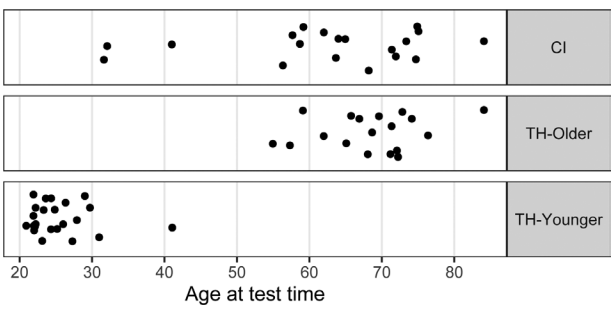

Fig. 1. Ages of participants in 3 listener groups.

CI listeners all had at least 3.5 year's experience with their device (median experience: 10 years). There was a median of 25 years duration of deafness until first implantation among the CI group, which included 10 unilaterally and 12 bilaterally implanted individuals. Among the 10 unilateral users, 5 regularly used a hearing aid in the contralateral ear to manage moderate to profound hearing loss, and continued using the hearing aid during the experiment.

All participants spoke English fluently, and none had acquired a hearing or language disorder before acquiring spoken language in childhood. Each participant provided informed consent to a procedure that was approved by the Institutional Review Board of the University of Minnesota.

Stimuli

Stimuli included 150 target sentences developed and recorded by Winn & Teece (2021) who provided a detailed description in their 2021 paper. Each sentence contained a target word early in the sentence (second, third, or fourth word) that was not predictable based on preceding words but was narrowly constrained based on subsequent words. The context would enable mental repair of the word in the case that it was masked during audio presentation. For example, "Please ____ the floor with this broom," where the target word is "sweep." The contextual constraint on the words was verified using a cloze test (where participants fill in the missing word in the written text of the sentence, and their responses are aggregated to verify the commonality of responses). The sentences were originally spoken slowly and clearly by an audiologist (the author) with explicit effort to facilitate high intelligibility, with the best recording of each sentence used for testing. The sentences were divided into five lists of 30, equalized for word length and position of the target word, with all sentenced equalized for root-mean-square amplitude. All of the sentence stimuli are available on the Open Science Framework website at the following URL: https://osf.io/ctnrj/.

Stimulus Types • There are two versions of each sentence, illustrated in Figure 2. The "Intact" version was the full utterance with all words spoken naturally. In the "Masked/Repair" version, the target word was replaced with noise that was matched in duration and intensity, whose frequency spectrum matched the long-term spectrum of the entire stimulus corpus.

Digit Triplets • Digits (excluding the bisyllabic number seven) were recorded by the same talker who produced the target sentences. Each digit was padded with silence at the offset to result in an audio file that was 2/3 sec long, so that there were always 2 sec separating the end of the sentence from the verbal response prompt that followed the digits.

Listening Conditions • Each sentence was followed by either 2 sec of silence or a digit triplet that lasted 2 sec. The listener was instructed to either ignore the digits or to repeat them along with the sentence; instructions applied to an entire block of 25 sentence stimuli. All stimuli were presented in quiet. Participants were always aware of the listening condition (digits-ignored, digits-repeated, or no digits), as they were blocked together and described to the participant before each block.

Procedure

Participants completed a sentence repetition task organized into blocks of 25 stimuli each. The blocks contained a random mix of intact and masked versions of the target sentences,

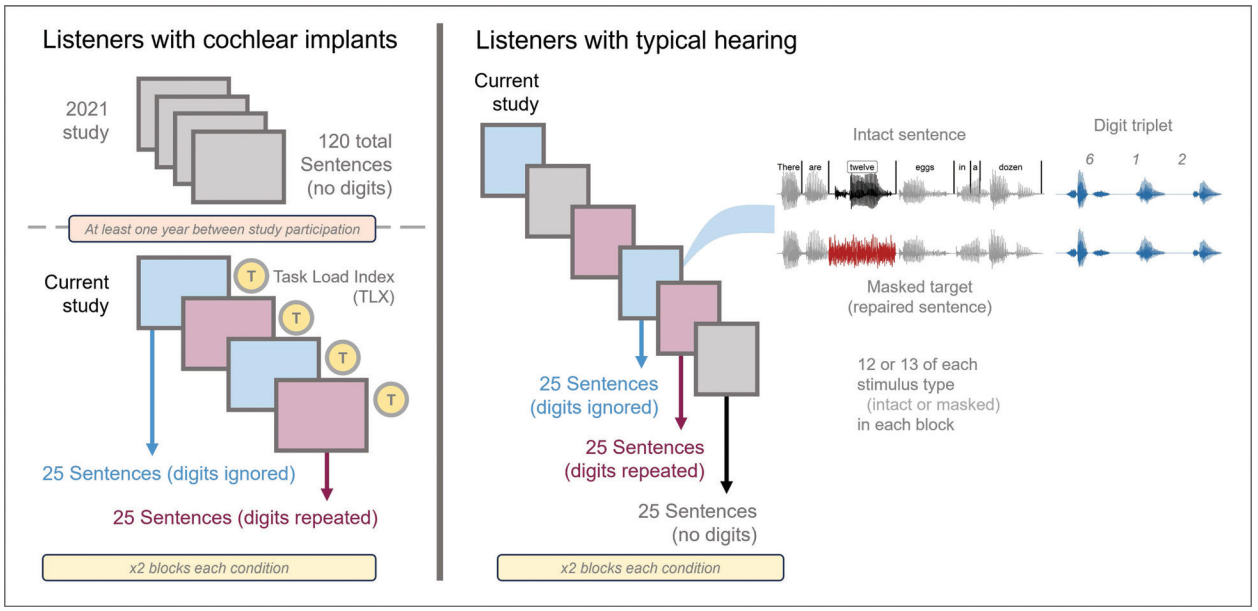

Fig. 2. Diagram of testing blocks in this study. Listeners with typical hearing heard all three stimulus types in the same session. CI listeners underwent testing for the stimuli followed by silence. At least 1 yr elapsed before completing the conditions with the sentences followed by digits. CI listeners also completed a NASA task load index questionnaire following each block of stimuli with digits. CI indicates cochlear implant.

and no sentence was repeated for any participant. They were instructed to always repeat the full sentence, and to guess if they were unsure (specifically: "repeat what you thought the talker said").

Each TH participant heard 25 sentences for both of the unique stimulus types (intact/masked) for each listening condition (most CI listeners heard 40 in each condition, in an attempt to prioritize stability of their data). Blocks were presented in randomized order to mitigate the effect of total testing time on each listening condition. Each list began with an intact sentence, followed by a random ordering of stimulus types, with no more than three consecutive trials of the same type. The presentation of lists was rotated and counterbalanced across listeners, and the type of stimulus for each item was rotated for each listener, except for the first trial in each list, which was always an intact sentence. Figure 2 illustrates the sequence of testing blocks completed by the participant groups.

In the CI group, 17 of the 22 participants had already completed the "no-digits" condition for a previous experiment, so their data for that condition were inherited for the analysis in the present study, and they only completed conditions involving digits after the sentence. For all of them, more than a year had passed because participating in the first condition, mitigating concern about remembering specific sentences. For all of these listeners, the no-digits condition was not repeated, in an attempt to reduce the overall number of test stimuli to mitigate listening fatigue (because this was a study of momentary effort rather than fatigue). The remaining 5 CI listeners followed the protocol as described earlier, where all three conditions were presented in the same test session.

After each block of sentences, CI participants completed the National Aeronautics and Space Administration (NASA) task load index (TLX), which is a multidimensional scale of subjective effort. The TLX includes dimensions for effort, mental demand, frustration, physical demand, and temporal demand. Although this questionnaire is not expected to be sensitive to the timing aspects of listening effort, it was used to gauge whether subjective perception of effort would change as a result of the overall listening condition.

During the experiment, participants sat in a chair with their forehead position stabilized by a cushioned bar. No chinrest was used, so that participants could comfortably move their jaw for speaking. Participants visually fixated on a red cross in the middle of a medium-dark gray background on a computer screen that was 50 cm away. Lighting in the testing room was kept constant. Each trial was initiated by the experimenter with at least 5 sec because the previous trial ended. The participant heard two beeps when the trial was about to begin. Two seconds after the beeps, the sentence was played at 65 dBA through a single loudspeaker in front of the listener. Two seconds after the sentence, the red cross turned green, which was the cue for the listener to give their response. If the testing block included digit triplets after the sentence, they were played during that 2-sec retention interval. The participants' verbal responses were scored on paper and also audio recorded for later inspection. The participant's eye position and pupil size were recorded by an SR Research Eyelink 1000 Plus eye tracker recording at 1000 Hz sampling rate, tracking pupil diameter in the remote-tracking mode, using the desktop-mounted 25 mm camera lens.

Analysis of Behavioral Data

Sentence Repetition • Repetition accuracy was scored in real time by an experimenter and documented for analysis. For stimuli where the target was replaced by noise, if the substituted word was not semantically coherent with the stimulus, it was counted as an error. If the participant's guess at the word that was replaced by noise was not the "intact" version of the word but still made sense (e.g., "The player shot the soccer ball into the goal" instead of "The player kicked the soccer ball into the

goal"), it was counted as correct. For all stimuli, we also tracked the presence of errors away from the target word, and whether participant responses were linguistically coherent. An example of an incoherent response would be "The plant hit the soccer ball with the door" (see Winn & Teece 2021 for further discussion of incoherent responses).

The main goal of the scoring system was to track whether the participant response had any error, rather than counting the number of word-specific errors within the response. This approach was taken specifically because the words in these high-context sentences were not independent; multiple errors within a sentence were not a conclusive sign that multiple words were misperceived. For example, misperception of a word might result from the listener trying to create coherence with an earlier word that was misperceived, and participants tend to produce these secondary errors (see Winn & Teece 2021; Gianakas et al. 2022 for evidence of this effect in both forward- and backward-direction within the sentence, and suggestion that a secondary error tends to reduce effort because it promotes coherence).

Sentence repetition scores were analyzed using a series of statistical models that estimated various outcome measures including (1) the presence of any error within the response, (2) errors specifically on the target word, (3) errors on words other than the target word, and (4) whether the response was incoherent. Errors were estimated on a per-trial level using a binomial (i.e., logistic) mixed-effects model that included fixed effects and interactions between the all three terms of listener group, stimulus type, and condition, as well as random intercepts and random effects of stimulus type and condition per listener (listener group was not a random effect for each listener because each listener contributed data to only a single group). The model formula was declared as follows:

$$\begin{aligned}&\text{glmer}(\text{any\_erro} \sim \text{Condition} \times \text{StimType} \times \text{Group}\\&\quad + (1 + \text{Condition} \times \text{StimType} \mid \text{Listener}))\end{aligned}$$

In this style of model, the glmer function is used to generalize the principles of a linear model to nonlinear data using the binomial linking function. The binomial outcome variable results in outcome (β/beta) coefficients that correspond to changes in the natural-log-odds of a change in response. This method of analysis accounts for the tendency to observe greater variability in the central range of performance, and magnifies differences at the extremes (i.e., the difference between 45% and 55% is treated as smaller than the difference between 85% and 95%), consistent with the distribution of this outcome. The model for estimating the presence of an error on words other than the target had the same structure as the model for any errors.

Models for true target errors and for incoherent responses were restricted only to CI listeners because there were not enough of these responses in the other groups (models included estimates for values at or close to zero, resulting in implausibly high- or low-β estimates). In the case of intact sentences, the target word was defined as the word that would have been masked by noise in the alternate version of the stimulus, to facilitate fair comparison across stimulus types. These models included fixed effects of listening condition and stimulus type, as well as correlated random intercepts and random slopes and interactions for listening condition and stimulus type, using the following formulae:

$$\begin{aligned}&\text{glmer}(\text{target\_error} \sim \text{Condition} \times \text{StimType}\\&\quad + (1 + \text{Condition} \times \text{StimType} \mid \text{Listener}))\\&\text{glmer}(\text{incoherent} \sim \text{Condition} \times \text{StimType}\\&\quad + (1 + \text{Condition} \times \text{StimType} \mid \text{Listener}))\end{aligned}$$

For the digits-repeated listening condition, the number of digits correctly repeated was estimated in a linear mixed-effects model (because it was not restricted to a binomial 0 or 1) using a fixed effects and interactions of listener group and stimulus type, as well as random intercepts and random effect for stimulus type for listeners, using the following formula:

$$\begin{aligned}&\text{lmer}(\text{num\_digit\_errors} \sim \text{Group} \times \text{StimType}\\&\quad + (1 + \text{StimType} \mid \text{Listener}))\end{aligned}$$

For all models, when a specific comparison was not available in the original model because both sides of the comparison were deviations from the default (and therefore not directly compared with each other), comparisons were obtained by rotating the same model with the default reassigned, rather than running a formal post-hoc model limited to the specific comparison of interest.

Subjective Effort Analysis • NASA-TLX subjective scales of effort range from 0 to 20; responses were fit with a linear mixed-effects model that included fixed effects of condition (digits-ignored or digits-repeated) and listener group, along with a random intercept and random slope for condition (Plus Random Condition × Interaction Effect) for each participant. This model could not include a term for stimulus type, as the effort ratings were solicited only at the ends of blocks that contained a random mix of both stimulus types. The formula was as follows:

$$\begin{aligned}&\text{lmer}(\text{rating} \sim \text{condition} \times \text{tlx\_dimension}\\&\quad + (1 + \text{Condition} \times \text{tlx\_dimension} \mid \text{Listener}))\end{aligned}$$

Analysis of Pupillometry Data

Pupil Data Preprocessing • Pupil data were preprocessed in the style described by Winn and Teece (2022). Blinks were detected as a decrease in pupil size to 0 pixels, and then the stretch of time corresponding to the blink was expanded backward by 80 msec and forward by 120 msec to account for the partial occlusion of the pupil by the eyelids during blinks. The blink was linearly interpolated, and then the signal was low-pass filtered at 5 Hz using a fourth-order Butterworth filter and then downsampled to 25 Hz. The baseline pupil size was calculated as the mean pupil size in the time spanning 500 msec before stimulus onset to 500 msec after sentence onset, and each pupil size data point in the trial was expressed as a proportional difference from the trial-level baseline.

Trials were discarded if 30% or more data points were missing between the start of the baseline to 3 sec past the onset of the stimulus. For all three listener groups, less than 3% of trials overall were dropped because of missing data. However, 1 CI listener, 1 older TH listener, and 2 younger TH listeners had

more than 10% of data dropped for this reason. Other outliers/contaminations were automatically detected through an algorithm that accumulated multiple "flags" to detect some of the artifacts discussed previously by Winn et al. (2018) and in greater depth by Mathôt et al. (2018) and Steinhauer et al. (2022), including: high-intensity pupil size oscillations (hippus) during baseline; baselines that had anomalous deviation (greater than 16% change) from both the previous and the next baseline; significant slope of change in pupil size during the baseline (more than 2.5 SD difference from the rest of trials); or a significant negative dilation immediately after the stimulus onset. More than two flags resulted in the entire trial being dropped. A total of 321 trials were dropped out of 8910 that were recorded, resulting in a drop rate of 3.6%. If any participant had fewer than 13 trials remaining in any single listening condition following outlier detection, that participant's entire dataset was dropped; 2 younger TH listeners were excluded for this reason.

Pupil Data Analysis • Pupil data were modeled using generalized additive mixed-effects models (GAMMs), which have been used in previous studies of time-series changes in pupil dilation (Poretta & Tucker 2019; Van Rij et al. 2019; Pandža et al. 2020), and eye-tracking data (Cychosz et al. 2023). GAMMs are similar to traditional generalized linear models, but the linear predictor partly depends on smoothing functions consisting of a combination of Gaussian (or other) basis functions that combine to form the nonlinear shape of the modeled data. GAMMs offer some distinct advantages over reporting mean or peak pupil dilation in that they account for the entire time course of the data, which is a key focus of the present experiment. In addition, GAMMs provide estimates of the timeline of differences rather than the mere existence of differences, as one might obtain from polynomial growth curve analysis. Statistical differences are identified when the 95% confidence interval of the estimated time-series data does not overlap with the estimate of the comparison series.

Another major strength of GAMMs is that they account for autocorrelation (the relation of each timepoint to the previous one), which is a notoriously challenging aspect of time-series data (like pupil size) that can lead to inflated risk for type-I errors (Baayen et al. 2022). The problem is that the value of each data point is highly related to neighboring data points, which are not independent observations. By estimating the degree of autocorrelation, the influence of the underlying parameters can be estimated more conservatively. Following the methods of previous work (Sóskuthy 2017, 2021), a two-step process was conducted to address this issue. First, the autocorrelation value was extracted from the model and then fed into the next generation of the model as the parameter rho, with the same fixed and random-effect terms. This two-generation approach to modeling allows the autocorrelation correction factor to be informed by the same parameters used in the prevailing model, but with the assumption that the errors for adjacent observations (i.e., time points of pupil data) are correlated.

All pupillometry data models were created in the R software using the bam function in the mgcv package (version 1.9; Wood 2017), with model evaluation done using the itsadug package (version 2.4.1; van Rij et al. 2022). Each model included parametric effects of stimulus type, a smooth interaction between time and stimulus type, as well as with random time smooths for each listener that interacted with stimulus type, as reflected in the model formula later.

$$\begin{aligned}&\text{bam}(\text{pupi\_dilation} \sim \text{StimType} + \\ &\text{s}(\text{time},\ \text{by} = \text{StimType},\ \text{k} = 26) + \\ &\text{s}(\text{time},\ \text{Listener},\ \text{by} = \text{StimType},\ \text{bs} = \text{fs}, \text{m} = 1), \\ &\text{family} = \text{scat},\ \text{method} = \text{fREML}, \text{discrete} = \text{TRUE}, \\ &\text{AR.start} = \text{start\_of\_trial},\ \text{rho} \\ &= \text{rho\_from\_no\_ac\_model})\end{aligned}$$

Note that in this model, the rho (autocorrelation control) parameter was calculated using the previous generation of the model that did not account for autocorrelation. The parameter k indicates the number of knots in the composite estimation of the data shape. The line that includes bs = 'fs' is the declaration of the random time smooths per listener, varying across stimulus types.

A follow-up model estimated the subtracted differences between curves for the intact versus repair-type stimuli directly, so that the cost of the mental repair process could be compared across groups and listening conditions. This model contained fixed parametric effects of listener group and listening condition and their interaction, along with separate time smooths for each listening condition, with random time smooths for each listening condition for each listener.

$$\begin{aligned}&\text{bam}(\text{diff\_curve} \sim \text{group\_x\_Condition} + \\ &\text{s}(\text{time},\ \text{by} = \text{group\_x\_Condition},\ \text{k} = 26) + \\ &\text{s}(\text{time},\ \text{Listener},\ \text{by} = \text{Condition},\ \text{bs} = \text{fs},\ \text{m} = 1), \\ &\text{AR.start} = \text{start\_of\_trial},\ \text{rho} \\ &= \text{rho\_from\_no\_ac\_diffcurve\_model})\end{aligned}$$

RESULTS

Sentence Repetition Errors

Errors Anywhere in the Response • Sentence repetition performance is displayed in Figure 3, broken down into a variety of error patterns. The CI group had statistically more errors than the older TH group ($\beta = 1.9$, $z = 4.77$, $p < 0.001$) and the younger TH group ($\beta = 1.6$, $z = 4.72$, $p < 0.001$). This effect of listener group was not statistically different in any of the listening conditions ($|z| < 0.9$ and $p > 0.38$ for All Condition × Group Interactions). The need to mentally repair target words had a substantial impact for CI listeners (red bars in Fig. 3A), leading to a roughly 17 percentage-point increase in mistaken trials compared with intact stimuli that was statistically detectable ($\beta = 1.07$, $z = 6.9$, $p < 0.001$). This effect for CI listeners was consistent across all listening conditions, as there was no significant interaction of stimulus type with the listening condition ($|z| < 0.7$, $p > 0.48$ for both Listening Condition × Stimulus-Type Interaction Terms).

In contrast to the CI group, the effect of stimulus type was significantly reduced (i.e., less detrimental) for the younger TH group, almost completely counteracting the effect observed in the CI group (group interaction $\beta = -1.01$, $z = -3.09$, $p = 0.002$). The impact of stimulus type on sentence repetition errors for the older TH group and was intermediate to that of the other two groups; the difference between older TH and CI groups in terms of the effect of stimulus type was close to but did not surpass

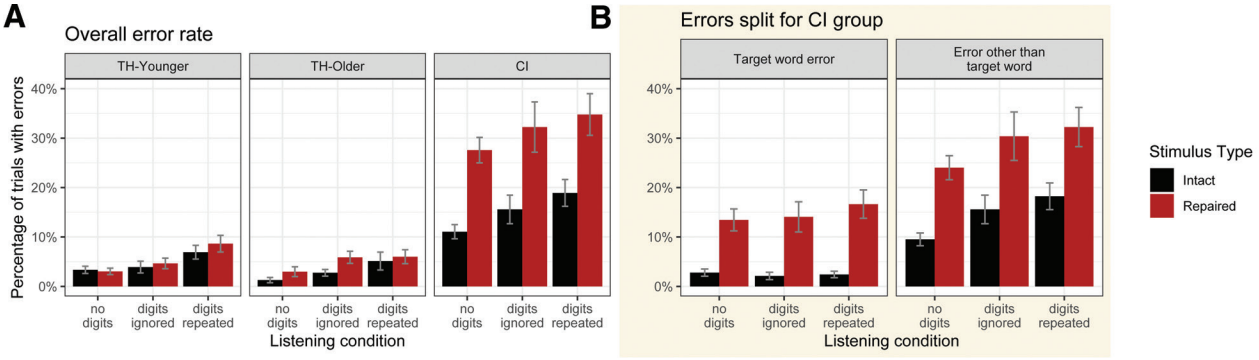

Fig. 3. Intelligibility results, showing the percentage of trials that contain various types of errors during sentence repetition. A (left), The proportion of trials with at least one error in the sentence. B (right), Breaks down these errors by the CI group into errors on the target word (the word that was masked and intended to be repaired in the repair condition, and its counterpart in the intact condition), vs. the presence of any errors elsewhere in the sentence. CI indicates cochlear implant.

the conventional threshold for statistical detection ($\beta = -0.67$, $z = -1.9$, $p = 0.058$). The Group × Stimulus Type Interactions for both TH groups were not statistically different across any of the listening conditions ($|z| < 0.79$ and $p > 0.43$ for all Group × Stimulus-Type × Condition Interactions).

On its own, the listening condition (i.e., digits after the sentence) generally had no substantial effect on the rate of sentence errors, with one exception; CI listeners made more errors repeating sentences that were followed by digits that were recalled compared with sentences followed by no digits ($\beta = 0.56$, $z = 2.86$, $p = 0.004$). When there were digits presented after the sentence, the instructions to repeat rather than ignore the digits resulted in an increase in errors that were not statistically significant ($\beta = 0.32$, $z = 1.6$, $p = 0.11$). A simplified view of all of the statistical effects described earlier is available in Supplemental Digital Content 1, http://links.lww.com/EANDH/B431, where estimated marginal means of all model terms are visualized along with a transformation of log-odds into proportional data.

Errors on the Target Word Versus Elsewhere in the Sentence • The right panel of Figure 3 splits sentence repetition errors among the CI group into errors on the target word versus words elsewhere in the sentence. These listeners were more likely to make errors on the target word when it was masked by noise than when it was not ($\beta = 2.42$, $z = 5.08$, $p < 0.001$), despite the highly constraining contextual words that were used by the other listener groups to infer those target words successfully. The need to repair a word in the stimulus also significantly increased the likelihood of errors elsewhere in the sentence for CI listeners ($\beta = 0.95$, $z = 5.99$, $p < 0.001$).

CI listeners consistently made more errors on words other than the target word compared with the older TH listener group ($\beta = 1.92$, $z = 4.73$, $p < 0.001$) and the younger TH group ($\beta = 1.54$, $z = 4.56$, $p < 0.001$). The interacting effect of stimulus type was significantly reduced (less detrimental) for the younger TH listeners ($\beta = -0.92$, $z = -2.78$, $p = 0.005$), but for the older TH listeners it was reduced by an amount that was not statistically different from the CI group ($\beta = -0.18$, $z = -0.46$, $p = 0.65$).

The Impact of Target Word Performance on Success Elsewhere in the Sentence

For CI listeners, success on the target word itself had downstream effects on words elsewhere in the stimulus. Although trials with error on the target word were relatively rare overall (2 out of 40 trials for intact sentences, and 6 out of 40 trials for sentences with the target masked), the presence of those target errors significantly increased the likelihood of making an error elsewhere in the sentence ($\beta = 2.68$, $z = 7.88$, $p < 0.001$). Supplemental Digital Content 2, http://links.lww.com/EANDH/B432, illustrates this effect, showing a roughly 40 to 60% point increase in non-target word error rate when target words were mistaken. This effect was not statistically different across the three listening conditions nor across the two stimulus types ($\beta = 0.14$, $z = 0.42$, $p = 0.68$). This pattern replicates and extends previous work showing the presence of an error early in a sentence is associated with a large increase in the rate of errors later in the sentence (Gianakas et al. 2022). Even when accounting for target word correctness, the status of the target word as masked (i.e., the need for repair) still resulted in a statistically higher likelihood of making errors on non-target words ($\beta = 0.6$, $z = 6.84$, $p < 0.001$), suggesting that even when repair was successful, there was a downstream cost that could not be explained by the absence of the contextual information that would have been provided by the target word itself.

Incoherent Responses Among CI Listeners

Figure 4 shows that CI listeners were more likely to give incoherent responses when the stimulus required mental repair. There was a roughly 7% increased risk of incoherent perceptions resulting from the need to mentally repair a sentence, which was statistically significant ($\beta = 2.27$, $z = 3.14$, $p = 0.002$). There was a higher rate of incoherent responses for stimuli where digits were recalled compared with when they were ignored, although this increase did not reach conventional threshold for statistical significance ($\beta = 1.48$, $z = 1.77$, $p = 0.076$).

Errors on Digits After the Sentence

The main novel outcome measure for this study was the performance for repeating digit triplets following the sentence (Fig. 5). For the older TH group, digit errors were more likely when the preceding sentence demanded mental repair ($\beta = 0.11$, $t = 2.4$, $p = 0.018$). The same effect was observed for the CI listeners, and was even larger (group interaction $\beta = 0.13$, $t = 2.12$, $p = 0.037$). Conversely, digit performance for the younger TH group was not affected by mental repair, with the rate of

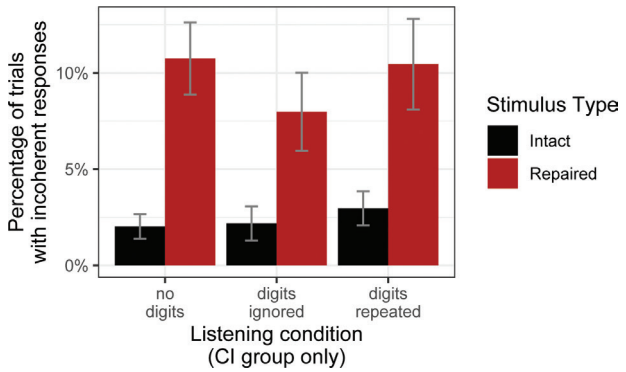

Fig. 4. Rate of linguistically incoherent responses in the CI group split by listening condition. CI indicates cochlear implant.

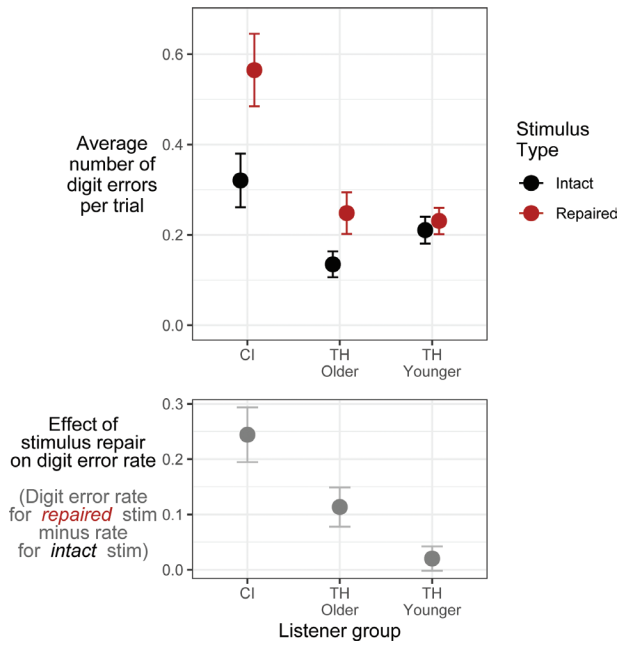

Fig. 5. Downstream errors on digit triplets following sentences. CI indicates cochlear implant.

digit errors increasing only by an average of only 0.02 digits per trial when the stimulus demanded mental repair, which was not statistically different from zero ($\beta = 0.02$, $t = 0.58$, $p = 0.56$). The older group with TH had performance on the digits that were statistically indistinguishable from the performance of the younger TH group, both for the digits following intact sentences ($\beta = -0.07$, $t = -1.16$, $p = 0.249$) and for digits following sentences that demanded mental repair ($\beta = 0.09$, $t = 1.45$, $p = 0.151$). These results collectively suggest a downstream cost of mental repair for older listeners, which is exacerbated when the listener uses a CI.

Although not shown in Figure 5, there were reduced digit errors on the first digit (suggesting a primacy benefit) and reduced errors on the final digit (suggesting a recency benefit); detailed analysis of this pattern is available in Supplemental Digital Content 3, http://links.lww.com/EANDH/B433.

The rate of downstream digit errors for CI listeners was further broken down into specific stimulus-response patterns in Figure 6 (with specific statistical comparisons available in Supplemental Digital Content 3, http://links.lww.com/EANDH/B433). This analysis shows that the mere need for mental repair in the sentence had only a modest effect on downstream digit errors, and that effect slightly increased when there was a genuine error that was still coherent with the meaning of the sentence. However, a much larger effect emerged when the listener's response was incoherent (i.e., not successfully repaired into a meaningful utterance); in that situation, the error rate on the digits more than doubled compared with when the sentence was repaired coherently, and increased more than fourfold compared with the performance following correctly-perceived intact sentences. This overpowering effect of response coherence is consistent with results of the previous studies related to the present work (Winn & Teece 2021, 2022), where response coherence was the largest effect of any that was evaluated.

Subjective Report of Effort

CI listeners' subjective reporting of effort is displayed in Figure 7, split by the separate dimensions of effort named by the TLX. In each dimension, the cognitive load was reported to be higher following blocks of trials where post-sentence digit triplets were repeated compared with when the digits were ignored. The smallest impact of listening condition was for the dimension Temporal Demand (the model default), where the effect of listening condition was still statistically detectable ($\beta = 1.6$, $t = 3.04$, $p = 0.005$). The impact of listening condition was statistically greater for the dimensions Effort ($\beta = 3.47$, $t = 4.17$, $p < 0.001$) and Mental Demand ($\beta = 2.64$, $t = 3.3$, $p = 0.003$), and the impact of listening condition was not statistically differentiable across the other three TLX dimensions.

Pupil Dilation

Responses to Intact Stimuli Versus Stimuli Demanding Mental Repair • Patterns of pupil dilation in each condition for all listener groups are illustrated in Figure 8A. For all listener groups, the need to mentally repair masked words in a sentence resulted in substantial increase in pupil dilation compared with sentences that were presented intact. Stretches of time where the curves were different are indicated by a gold bar at the bottom of each panel in Figure 8A.

Comparing the Cost of the Mental Repair Process Across Listening Conditions • The effect of mentally repairing a masked word in the sentence is illustrated directly in Figure 8B, with each line reflecting the differences between curves for each panel in Figure 8A. This analysis compares the cost of mental repair across the listening conditions. For all listener groups, the repair cost was statistically larger when the sentence was followed by silence rather than a digit triplet (the blue and pink lines are lower than the black line). In the plot, these differences are illustrated by alternating black/blue and black/pink dashed lines denoting the stretch of time where the difference between curves was found to be statistically detectable. For all three groups, there was no statistical difference in the repair cost between the digits-ignored and digits-repeated conditions.

Effort Reallocation During the Sentence • A novel question in this study was whether overall patterns of pupil dilation would change when the sentence was accompanied by silence, the presence of digits, and the need to recall the digits. These data are visualized in Figure 9, showing changes in pupil size averaged across both stimulus types (repair/intact) to more

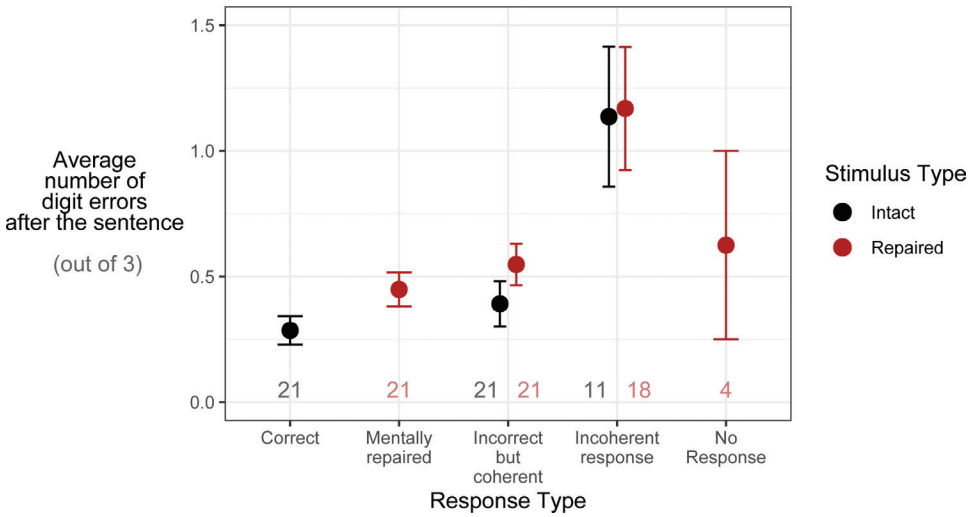

Fig. 6. Downstream errors on digit triplets following sentences, broken down by stimulus-response patterns. Numbers above each response category indicate the numbers of unique participants who contributed data to that category (i.e., not all listeners gave incoherent responses, so that number is lower than the total number of participants).

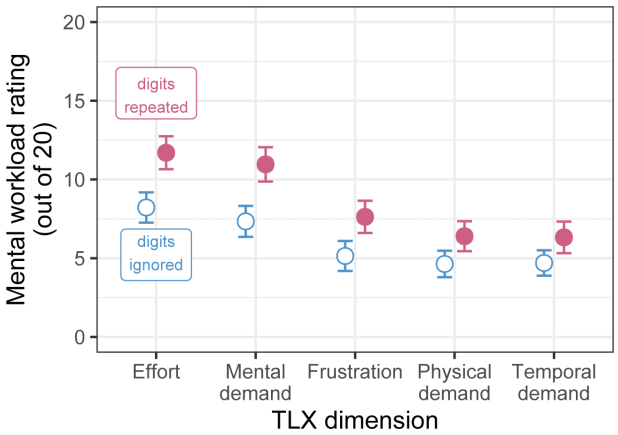

Fig. 7. Subjective ratings of cognitive load reported by CI participants for the listening conditions where digits followed each sentence (many listeners completed the no-digits condition before the TLX was adopted into the protocol, so that condition is not displayed in the plot). CI indicates cochlear implant; GAMM, generalized additive mixed-effects model; TLX, task load index.

easily view the effect of listening condition on its own. Data for the no-digits condition for CI listeners were not estimated in this analysis due to the large gap between testing times for this specific component of the experiment (i.e., the sessions were far enough apart that there could be concerns about comparisons of data that reflect physiological engagement).

A GAMM was constructed to estimate pupil size in Figure 9 as a product of parametric terms for listening condition and group, as well as an ordered-factor difference-smooth term for stimulus type so that the effect of condition could be estimated separately from the main effect of interest. Random effects included time-smooth interactions with stimulus type for each listener, with correction for autocorrelation as described earlier. Results of the model are visualized in the lower panel of Figure 9, showing a difference in the timing of pupil dilation depending on the instructions in the listening condition. The younger TH listeners showed increased pupil size for the digits-ignored condition in the early part of the trial (specifically between −1.1 and 0.68 sec relative to sentence offset), and then this relationship flipped, with greater pupil size for the digits-repeated condition in the later part of the trial (from 1.06 sec relative to sentence offset, all the way through the end of the analysis window). The other two listener groups showed only half of this pattern, lacking any statistical difference between listening conditions until 1.5 and 1.88 sec after sentence offset for the TH-older and CI groups, respectively. In other words, only the younger TH listeners showed "front-loading" of increased pupil size for the digits-ignored condition. This result replicates earlier findings by Winn and Moore (2018), who observed differences in the timing of pupil dilation based on task instructions for TH but not for CI listeners.

DISCUSSION

The Main Hypotheses

Results verified hypothesis No. 1, which was the most important novel aspect of the present study: when sentences demanded mental repair, CI listeners were more likely to make mistakes on the digits that immediately followed the sentence (Figs. 5 and 6). This was also true to a slightly lesser extent for older TH listeners. These results suggest that carryover effort from the mental repair process might last long enough to potentially jeopardize ongoing conversation, which would have more complex demands than the simple digit triplet task used in the present study. This effect was driven largely by incoherent responses. In contrast, results for younger TH listeners failed to show any impact of mental repair on downstream listening, suggesting that the process of mental repair can be quick enough to spare these listeners from any disruptive effects in this task. It is possible that the younger TH listeners capitalized on the clarity of the supporting contextual information after the missing target word to recover more quickly.

Hunter (2021) suggested that easier sentence processing facilitated by contextual cues would rescue the listener from the competing demands of a secondary task, so long

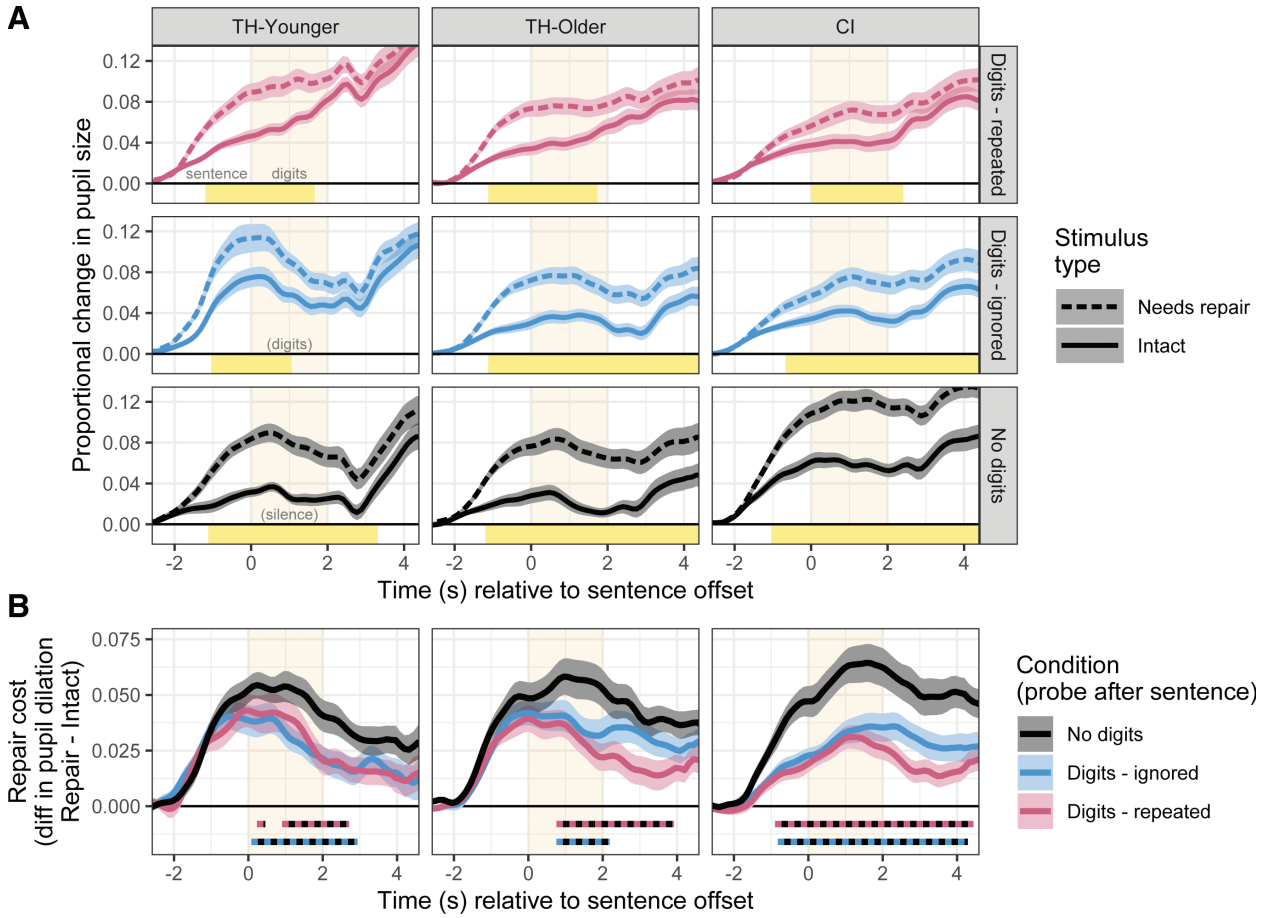

Fig. 8. Pupil dilation responses for the two stimulus types (intact or repaired) in each of three conditions. The light tan boxes indicate the 2 sec after the offset of the sentence (which would either be quiet or contain the digit triplet). A, The solid lines show responses for intact stimuli, while dashed lines reflect responses for stimuli with masked words, demanding mental repair. Yellow regions along the bottom of each panel denote stretches of time when those lines were statistically different in the GAMM. B (bottom), The lines reflect the linear differences between each corresponding line in the top section, operationalizing the cost of mentally repairing the sentence. In those panels, dotted lines reflect regions of statistical differences between the lines whose colors match the dots (e.g., black and blue dotted lines show intervals where the black line is different than the blue line). GAMM indicates generalized additive mixed-effects model.

as that secondary task was not overly demanding. The present study supports and expands this notion, showing that easier sentence processing (because there is no need to repair misperceptions) reduced errors on the secondary task of digit recognition after the sentence for CI listeners and older TH listeners.

When sentences needed to be mentally repaired, CI listeners were more likely to make errors elsewhere within the sentence (right panel of Fig. 3), supporting hypothesis No. 2. This was observed even when the target word was repaired correctly, suggesting an impact of the mental repair process itself. However, non-target errors were driven even more strongly by an actual error on the target, opening up various potential interpretations. For example, non-target errors might result from the listener trying to coerce coherence with the mistaken target (consistent with Winn & Teece 2021), or result from the relatively reduced amount of contextual information that would have been provided by a correctly-perceived target word. Regardless, this pattern supports the notion that words in sentences are not perceived independently.

Mental repair of missing words also led to a greater likelihood of responses that were linguistically incoherent (Fig. 4). Although incoherent responses were relatively rare, they are important because they have been found in two previous studies to be a larger driver of listening effort than any experimenter-controlled variable (Winn & Teece 2021, 2022). The present results suggest that these incoherent perceptions also have downstream consequences for later listening, as shown by reduced performance for the digits after sentences that were perceived incoherently. It is possible that mental repair demands cognitive resources that might otherwise be used to generate feasible alternative perceptions.

The burden of mental repair was observed as elevated pupil dilation (Fig. 8A) even when sentences were ultimately reported correctly, validating hypothesis No. 3. This supports the notion that successful sentence repetition by itself does not protect against elevated listening effort, which is likely better explained by the process of generating and evaluating candidate perceptions. For the TH listener groups, there was an unexpected increase in repair cost for the no-digits condition. Although this might result from the greater opportunity to devote attention to the sentence (as opposed to splitting resources across the sentence and subsequent digits), it might have also resulted from smaller pupil dilation for the intact stimuli in this condition, compared with the same stimuli in the conditions with post-sentence digits. The presence of digits

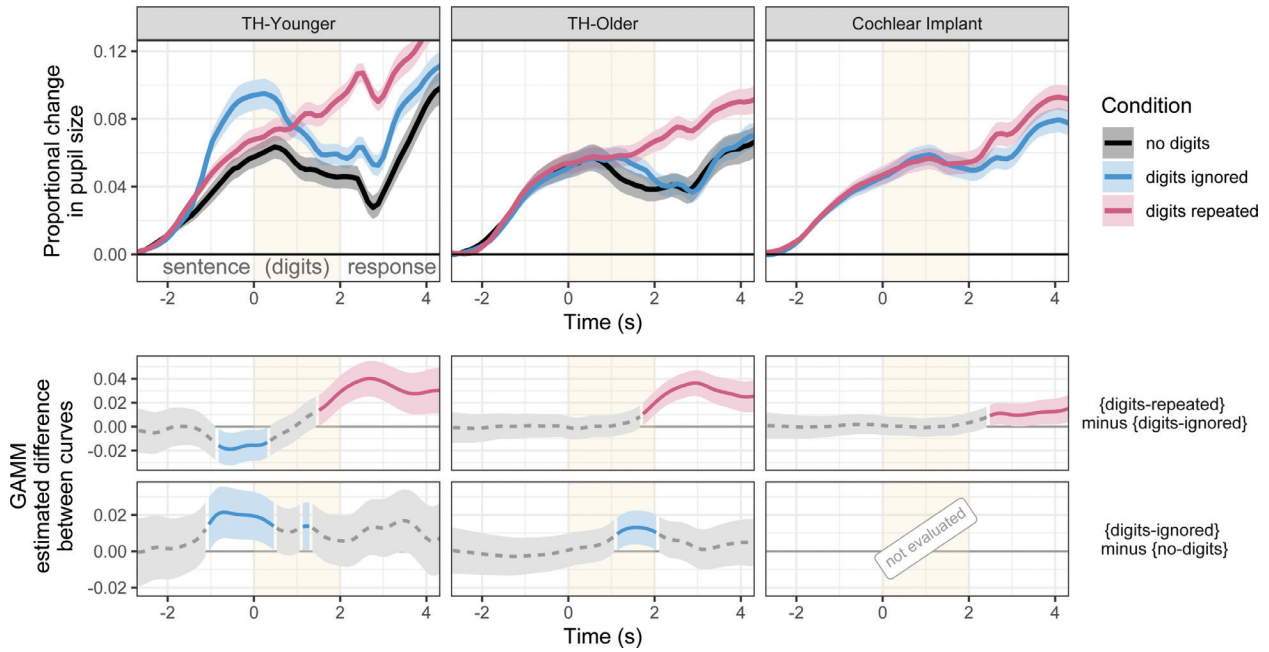

Fig. 9. Pupil dilation responses for the three listening conditions. The light tan boxes indicate the 2 sec after the offset of the sentence (which would either be quiet or contain the digit triplet). The upper row represents the data, and the bottom rows show the extracted difference curves from the generalized additive mixed-effects model. Stretches of time where the two curves were statistically difference (95% confidence interval of the difference-smooth curve excluded zero) are indicated by the color of the line that was greater in magnitude. Gray dashed stretches of the line indicate regions of non-significance.

at all in the stimulus—even if ignored—might have elevated arousal to an extent that reduced the differences in responses to the two stimulus types.

Hypothesis No. 4 was not verified, as sentence repetition accuracy was not substantially affected by the anticipation of upcoming digit triplets after the sentence, regardless of whether the digits were ignored or repeated (Figs. 3 and 4). This result contrasts with results reported by Hunter (2021), where digits before the sentence impose measurable detrimental effects on sentence repetition. The competing demands of the sentences and pre-sentence digits in Hunter's (2018, 2021) studies were bi-directional, suggesting that one could not perfectly isolate the contributions of either stimulus.

Subjective reporting of effort suggested that the need to repeat digits after the sentence resulted in greater perceived effort across all dimensions of the NASA-TLX questionnaire, with the task most strongly affecting the dimensions of Effort and Mental Demand regardless of listening condition.

Speech Repetition Scores Do Not Reflect Carryover Effort

As a group, the CI listeners in the present study made more errors in repeating back sentences and also made more errors on the digits that followed the sentences. However, these two outcomes were not correlated within the CI group ($r^2 = 0.0005$, $p = 0.931$). Trials that contained a sentence repetition error were not consistently the trials that contained an error on the digit probe after the sentence, and the listeners who demonstrate lower repetition scores for the sentences overall were not the listeners who show greater tendency to make errors on the digits. Carryover effort therefore cannot be predicted purely on the basis of the repetition errors made on basic sentence stimuli.

The notion that carryover effects from one sentence to the next cannot be predicted from correct repetition of the individual components has been found in numerous previous studies, including those where the difficulty of continuous speech materials is driven by background noise (Nagaraj 2017) or by the talker's dysarthria (Hustad 2008). Furthermore, the cognitive factors that were associated with speech repetition accuracy in a study by Nagaraj (2017) were found to not add any explanatory power when modeling comprehension of continuous speech. Auditory distortions (e.g., noise, visual cues) known to impair repetition scores do not have an equivalent effect on discourse comprehension (Tye-Murray et al. 2008). Collectively, these studies demonstrate that the demands of multi-utterance listening cannot be understood on the basis of performance of single utterances; there is a need for direct testing, using methods designed specifically to be sensitive to carryover effects from one utterance to the next.

Performance scores for isolated word recognition in quiet and AzBio sentences do not have any consistent relationship with patient-reported activity limitations, social interactions, perceived emotional handicap, or self-esteem (Capretta & Moberly 2016). In addition, McRackan et al (2018) found essentially no correlation of hearing-related quality of life with repetition of isolated words in quiet ($r^2 = 0.05$), repetition of sentences in quiet ($r^2 = 0.06$), or sentences in noise ($r^2 = 0.07$). Although quality of life has a potential relationship with audiovisual speech perception and complex sentence recognition (Moberly et al. 2018), the basic task of speech repetition assessments falls short of capturing factors that drive perceived handicap among people who use CIs. These results highlight the need to evaluate effort directly, rather than hoping it will be revealed indirectly while pursuing some other measurement (c.f. Beechey 2022).

Designing Tasks Intentionally to Detect Effort

Reflecting on a combination of listening effort studies that measured EEG alpha suppression, Wisniewski et al. (2021) pointed to the oversimplification of the concept of effort as a contributor to a literature of seemingly ambivalent results that limit the progress of understanding effort. Calling for more careful critique of task design, they highlighted how quantification of effort for a task cannot be disentangled from the variety of different approaches that a person can take to complete the task. Furthermore, they reinforced the concept of listening effort as decomposable into a series of events that engage multiple components of effort, rejecting the attraction of a unidimensional construct. Their approach is especially amenable to understanding speech, which is a series of events that build over time in nonlinear ways. In addition to the temporal components, Strauss and Francis (2017) further decomposed effort into sources arising internally or externally. In the present study, internal attention was aimed mainly at resolving ambiguity in perception by making inferences about a missing piece of the signal (the effect of stimulus type), whereas the effect of listening condition involved suppressing attention to digit triplets after a target sentence, which by contrast would be deemed external attention. The signature of external attention was mainly a reallocation of pupil dilation that was observed in the younger TH group but not the other groups. Perhaps pupil dilation is sensitive to the aspect of listening relating to the uncertainty and timing of making decisions (Satterthwaite et al. 2007; Lempert et al. 2015), but less sensitive to other factors that make perception difficult, such as external drivers of effort.

Response Coherence Drives Effort More Than Repetition Accuracy

In recent studies, sentence repetition accuracy scores were found to be an inadequate reflection of the effort of sentence recognition (Winn & Teece 2021, 2022), with whole-utterance coherence being a more powerful driver of cognitive load. Consistent with that notion, the performance for post-sentence digit recall (i.e., the listening effort carryover) in the present study was virtually unaffected by the mere presence of a mistake in the sentence recall, but was affected to a greater extent by the need for mental repair (i.e., need to build coherence) and even more by the inability to resolve a coherent sensible meaning in the sentence (Fig. 6). The lack of impact of a mere repetition error on its own is consistent with Hunter's (2021) study, where secondary task measures (digit recall and reaction times) were facilitated by coherent sentence context regardless of sentence repetition accuracy scores.

The importance of linguistic coherence as a driver of listening effort across multiple studies implies that test stimuli ought to have the possibility of incoherence for this effect to emerge. Previous results support the utility of complex materials to detect effects that are overlooked with simpler stimuli (McRackan et al. 2018; Moberly et al. 2018). However, a simple carrier phrase (e.g., "say the word … ") can create sentence-length stimuli but without the key ingredient of building meaning out of coherence across words. Supporting this notion, Ryan et al. (2023) tested perception of individual target words embedded in that carrier phrase, finding decreased attentional engagement (theta oscillatory power) during the carrier phrase itself. The authors suggested that meaningful sentences would likely show more robust changes in measures of theta and alpha power because they would be more cognitively engaging. It is not clear whether an incoherent perception would elicit effort only if it is unexpected; experiment designs that feature all or mostly incoherent stimuli (c.f. Mechtenberg et al. 2023) might diminish this effect because the drive to create coherence might be suppressed. Despite the focus on semantic coherence and inference in the present study, there are still other factors that influence effort that can emerge with shorter stimuli like single words (Kuchinsky et al. 2013; Colby & McMurray 2021; McLaughlin et al. 2022). Therefore there is certainly room for a wide variety of auditory tasks to detect varying levels of hearing difficulty.

Effort Anticipation and the Allocation of Effort Across Time

Younger TH listeners in the present study showed patterns pupil dilation that indicate strategic timing of their auditory attention. Pupil dilation rose earlier for stimuli where the key information was earlier (where later information was ignored), and dilation grew at a slower rate when the key information was spread out over a longer stretch of time (when they needed to attend to a sentence and the following digits). Considering that the conditions were blocked separately, the slope of pupil dilation might reflect the listener's anticipation of how long the relevant stimulus would last. Conversely, the CI listeners and older TH listeners did not clearly show this pattern, having roughly similar responses to the digits-ignored and the digits-repeated conditions with respect to the growth of pupil dilation leading to the peak. These results replicate findings by Winn and Moore (2018) who used the same task instructions with a different set of stimuli. The present results also substantiate the recommendation by Wisniewski et al. (2021) to consider the timing of effortful events rather than collapsing them into a single dimension.

Variation in the conditional front-loading or pacing of effort was also observed by Svirsky et al. (2024). In their experiment, the pupil response during the presentation of a sentence grew at a slower rate when listeners were cued to anticipate two sequential sentences. This result suggested that listeners might reserve some capacity for later moments of listening so that cognitive resources are not exhausted during the first utterance. Notably, this result emerged only for those CI listeners whose speech recognition score was minimally affected by the presence of the second sentence in the series; CI listeners who showed a two-sentence decrement failed to show this effort allocation pattern. Gathering observations by Svirsky et al., Winn and Moore (2018), and the present study, it appears that listeners who excel at speech recognition (either younger TH listeners or better-performing CI listeners) show temporal allocation of listening effort that reflects the demands of the task and in-the-moment anticipation. The absence of this effect might be explained by a phenomenon observed by Zekveld et al. (2019) who found that the reliable effects of SNR/sentence intelligibility for pupil responses were entirely overridden by the presence of an extra memory load task during the experiment.

Supporting the notion that effort allocation reflects strategic use of cognitive resources, recent results by Johns et al. (2024) suggest that listening effort depends at least partly on the extent to which individuals have mobilized their attention in anticipation of the difficulty of the upcoming task. Vaden et al. (2022) suggest that listeners might modulate cognitive resources in

anticipation of needing them for an upcoming task, finding better performance when trials were preceded by greater activity in the cingulo-opercular frontal network. Consistent with this, Mechtenberg et al. (2023) found increased pupil dilation before predictably difficult trials, and Micula et al. (2022) found a tendency for greater pupil dilation during trials that were correctly recalled, suggesting that increased engagement in the moment preceding the stimulus might promote better perception. Distinct patterns can arise when examining responses to several sentences as well (Bönitz et al. 2021). Exploring the allocation of listening effort might continue to reveal additional insights not available when tracking the overall peak magnitude of pupil dilation.

Limitations on Data Collection

The present study had an imbalance of protocol across the groups, requiring some caution before truly solidifying conclusions about the effect of using a CI. Specifically, much of the CI data for the no-digits condition were collected more than a year before data for the other conditions, while the other listener groups heard all three conditions in the same session. Although this presents a limitation for some comparisons of condition effects across the groups—particularly those that depend on comparisons of the magnitude of dilation—there is no obvious reason to discount these data. Incidentally, the reduction of stimuli because of the inclusion of two rather than three conditions offered the advantage of potentially mitigating listening fatigue in the CI group, whereas the other groups did not show any obvious signs of fatigue while doing three conditions. The effects of fatigue cannot be known from this study because of the nature of the protocol, but intentional choices to reduce fatigue could be a way to protect the deterioration of the phasic event-level signatures of effort like those sought in the present study.

Translating the Results to Everyday Life

There is potential for applying these results to audiological counseling and assessment. Regular conversation partners of people who are deaf or hard-of-hearing can better ensure the intelligibility of their own speech (and perhaps reduce the effort of listening to their speech) by including pauses that allow the listener to repair any recent misperceptions and arrive at coherent understanding of the message. Wingfield et al. (1999) demonstrated the protective effect of strategic pauses for improving speech repetition accuracy, and there is potential to find protective effect against effort as well.

Audiological assessment stands to improve detection of hearing difficulty by introducing post-sentence probes to test the listener's reliance on an extra moment to solidify the perception (c.f. Gianakas et al. 2022) or to test whether the effort of auditory processing might last long enough to interfere with upcoming speech (as in the present study). The main added value of this approach is to detect the listener who appears to function successfully for single words or sentences, but does so using a listening strategy that would not enable perception of continuous speech. Even in absence of an immediate solution to carryover effort, there could be value in communicating this difficulty that might not be intuitive to communication partners who have TH. Ultimately these efforts can drive toward testing a wider range of abilities involved in speech perception, as described by Beechey (2022).

CONCLUSIONS

Mentally repairing misperceived words has an immediate cost of greater effort. For CI listeners, mental repair also leads to a downstream cost of being more likely to miss the next thing that would be heard. That carryover effect is not predictable based on repetition accuracy for the sentence but is substantially greater when the listener cannot arrive at a sensible perception from the previous sentence. These results imply that listeners with CIs are likely to benefit from the insertion of pauses in speech to resolve any lingering perceptual ambiguities. Younger TH listeners appear to reallocate effort dynamically in response to specific task demands, but that ability is not clearly observed in CI listeners.

ACKNOWLEDGMENTS

This research was supported by NIH R01DC017114 (M.B.W.). Data collection and participant recruitment were coordinated by Katherine Teece. Data collection was assisted by Emily Hugo and Justin Fleming. The experiment design was assisted by our late colleague Akira Omaki. All data from this study and code to run aggregation and plotting can be found at: https://osf.io/n4urd/. This paper appeared as a preprint at: https://psyarxiv.com/txm5q.

The University of Minnesota stands on Miní Sóta Makhóčhe, the homelands of the Dakhóta Oyáte.

This study was approved under the University of Minnesota Institutional Review Board study 4150.

The authors have no conflicts of interest to disclose.

Address for correspondence: Matthew B. Winn, 164 Pillsbury Dr SE Minneapolis, MN 55455. E-mail: mwinn@umn.edu

OPEN PRACTICES

This manuscript qualifies for an Open Data badge and an Open Materials badge. The materials and data have been made publically available at https://osf.io/n4urd/ and https://osf.io/ctnrj/. More information about the Open Practices Badges can be found at https://journals.lww.com/ear-hearing/pages/default.aspx.

Received August 25, 2023; accepted May 14, 2024; published online ahead of print June 18, 2024

REFERENCES

Baayen, R. H., Fasiolo, M., Wood, S., Chuang, Y. Y. (2022). A note on the modeling of the effects of experimental time in psycholinguistic experiments. *Mental Lexicon*, *17*, 178–212.

Balling, L. W., Morris, D. J., Tøndering, J. (2017). Investigating lexical competition and the cost of phonemic restoration. *J Acoust Soc Am*, *142*, 3603.

Beechey, T. (2022). Is speech intelligibility what speech intelligibility tests test? *J Acoust Soc Am*, *152*, 1573.

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

Capretta, N. R., & Moberly, A. C. (2016). Does quality of life depend on speech recognition performance for adult cochlear implant users? *Laryngoscope*, *126*, 699–706.

Colby, S., & McMurray, B. (2021). Cognitive and physiological measures of listening effort during degraded speech perception: Relating dual-task and pupillometry paradigms. *J Speech Lang Hear Res*, *64*, 3627–3652.

Crowson, M. G., Franck, K. H., Rosella, L. C., Chan, T. C. Y. (2021). Predicting depression from hearing loss using machine learning. *Ear Hear*, *42*, 982–989.

Cychosz, M., Mahr, T., Munson, B., Newman, R., Edwards, J. R. (2023). Preschoolers rely on rich speech representations to process variable speech. *Child Dev*, *94*, e197–e214.

Gianakas, S. P., Fitzgerald, M. B., Winn, M. B. (2022). Identifying listeners whose speech intelligibility depends on a quiet extra moment after a sentence. *J Speech Lang Hear Res*, *65*, 4852–4865.

Hughes, S. E., Hutchings, H. A., Rapport, F. L., McMahon, C. M., Boisvert, I. (2018). Social connectedness and perceived listening effort in adult cochlear implant users: A grounded theory to establish content validity for a new patient-reported outcome measure. *Ear Hear*, *39*, 922–934.

Hughes, S. E., Watkins, A., Rapport, F., Boisvert, I., McMahon, C. M., & Hutchings, H. A. (2021). Rasch analysis of the listening effort questionnaire-cochlear implant. *Ear Hear*, *42*, 1699–1711.

Hunter, C. R. (2021). Dual-task accuracy and response time index effects of spoken sentence predictability and cognitive load on listening effort. *Trends Hear*, *25*, 23312165211018092.

Hunter, C. R., & Pisoni, D. B. (2018). Extrinsic cognitive load impairs spoken word recognition in high- and low-predictability sentences. *Ear Hear*, *39*, 378–389.

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *J Speech Lang Hear Res*, *51*, 562–573.

Johns, M. A., Calloway, R., Decruy, L. P., Karunathilake, I. M. D., Anderson, S., Simon, J. Z., Kuchinsky, S. E. (2024). Attention mobilization as a modulator of listening effort: Evidence from pupillometry. *Trends Hear*, *28*, 23312165241245240.

Kramer, S. E., Kapteyn, T. S., Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *Int J Audiol*, *45*, 503–512.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Jr, Cute, S. L., Humes, L. E., Dubno, J. R., Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology*, *50*, 23–34.

Lempert, K. M., Chen, Y. L., Fleming, S. M. (2015). Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLoS One*, *10*, e0126588.

Mathôt, S., Fabius, J., Van Heusden, E., Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behav Res Methods*, *50*, 94–106.

McLaughlin, D. J., Zink, M. E., Gaunt, L., Spehar, B., Van Engen, K. J., Sommers, M. S., Peelle, J. E. (2022). Pupillometry reveals cognitive demands of lexical competition during spoken word recognition in young and older adults. *Psychon Bull Rev*, *29*, 268–280.

McRackan, T. R., Bauschard, M., Hatch, J. L., Franko-Tobin, E., Droghini, H. R., Nguyen, S. A., Dubno, J. R. (2018). Meta-analysis of quality-of-life improvement after cochlear implantation and associations with speech recognition abilities. *Laryngoscope*, *128*, 982–990.

McRackan, T. R., Velozo, C. A., Holcomb, M. A., Camposeo, E. L., Hatch, J. L., Meyer, T. A., Lambert, P. R., Melvin, C. L., Dubno, J. R. (2017). Use of adult patient focus groups to develop the initial item bank for a cochlear implant quality-of-life instrument. *JAMA Otolaryngol Head Neck Surg*, *143*, 975–982.

Mechtenberg, H., Giorio, C., Myers, E. B. (2023). Pupil dilation reflects perceptual priorities during a receptive speech task. *Ear Hear*, *45*, 425–440.

Micula, A., Rönnberg, J., Książek, P., Murmu Nielsen, R., Wendt, D., Fiedler, L., Ng, E. H. N. (2022). A Glimpse of memory through the eyes: Pupillary responses measured during encoding reflect the likelihood of subsequent memory recall in an auditory free recall test. *Trends Hear*, *26*, 23312165221130581.

Moberly, A. C., Harris, M. S., Boyce, L., Vasil, K., Wucinich, T., Pisoni, D. B., Baxter, J., Ray, C., Shafiro, V. (2018). Relating quality of life to outcomes and predictors in adult cochlear implant users: Are we measuring the right things? *Laryngoscope*, *128*, 959–966.

Nachtegaal, J., Festen, J. M., Kramer, S. E. (2012). Hearing ability in working life and its relationship with sick leave and self-reported work productivity. *Ear Hear*, *33*, 94–103.

Nachtegaal, J., Kuik, D. J., Anema, J. R., Goverts, S. T., Festen, J. M., Kramer, S. E. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. *Int J Audiol*, *48*, 684–691.

Nagaraj, N. K. (2017). Working memory and speech comprehension in older adults with hearing impairment. *J Speech Lang Hear Res*, *60*, 2949–2964.

O'Leary, R. M., Neukam, J., Hansen, T. A., Kinney, A. J., Capach, N., Svirsky, M. A., Wingfield, A. (2023). Strategic pauses relieve listeners from the effort of listening to fast speech: Data limited and resource limited processes in narrative recall by adult users of cochlear implants. *Trends Hear*, *27*, 23312165231203514.

Pandža, N., Phillips, I., Karuzis, V., O'Rourke, P., Kuchinsky, S. (2020). Neurostimulation and pupillometry: New directions for learning and research in applied linguistics. *Annu Rev Appl Linguist*, *40*, 56–77.

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

Poretta, V., & Tucker, B. (2019). Eyes wide open: Pupillary response to a foreign accent varying in intelligibility. *Front Commun*, *4*, 1–12. https://doi.org/10.3389/fcomm.2019.00008.

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

Ryan, D. B., Eckert, M. A., Sellers, E. W., Schairer, K. S., McBee, M. T., Ridley, E. A., Smith, S. L. (2023). Performance monitoring and cognitive inhibition during a speech-in-noise task in older listeners. *Semin Hear*, *44*, 124–139.

Satterthwaite, T. D., Green, L., Myerson, J., Parker, J., Ramaratnam, M., Buckner, R. L. (2007). Dissociable but inter-related systems of cognitive control and reward during decision making: Evidence from pupillometry and event-related fMRI. *Neuroimage*, *37*, 1017–1031.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: A practical introduction. *arXiv preprint arXiv:1703.05339*. https://doi.org/10.48550/arXiv.1703.05339

Steinhauer, S. R., Bradley, M. M., Siegle, G. J., Roecklein, K. A., Dix, A. (2022). Publication guidelines and recommendations for pupillary measurement in psychophysiological studies. *Psychophysiology*, *59*, e14035.

Strauss, D. J., & Francis, A. L. (2017). Toward a taxonomic model of attention in effortful listening. *Cogn Affect Behav Neurosci*, *17*, 809–825.

Svirsky, M. A., Neukam, J. D., Capach, N. H., Amichetti, N. M., Lavender, A., Wingfield, A. (2024). Communication under sharply degraded auditory input and the "2-sentence" problem. *Ear Hear* Advance online publication. https://doi.org/10.1097/AUD.0000000000001500.

Szostak, C. M., & Pitt, M. A. (2013). The prolonged influence of subsequent context on spoken word recognition. *Atten Percept Psychophys*, *75*, 1533–1546.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *Int J Audiol*, *47*(Suppl 2), S31–S37.

Vaden, K. I., Jr, Teubner-Rhodes, S., Ahlstrom, J. B., Dubno, J. R., Eckert, M. A. (2022). Evidence for cortical adjustments to perceptual decision criteria during word recognition in noise. *Neuroimage*, *253*, 119042.

van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., Wood, S. N. (2019). Analyzing the time course of pupillometric data. *Trends Hear*, *23*, 2331216519832483.

van Rij, J., Wieling, M., Baayen, R., van Rijn, H. (2022). *Itsadug: Interpreting time series and autocorrelated data using GAMMs*. R package version 2.4.1. https://cran.rproject.org/package=itsadug.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393.

Winn, M. B. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends Hear*, *20*, 233121651666972–233121651666917.

Winn, M. B., & Moore, A. N. (2018). Pupillometry reveals that context benefit in speech perception can be disrupted by later-occurring sounds, especially in listeners with cochlear implants. *Trends Hear*, *22*, 2331216518808962–2331216518808922.

Winn, M. B., & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends Hear*, *25*, 23312165211027688.

Winn, M. B., & Teece, K. H. (2022). Effortful listening despite correct responses: The cost of mental repair in sentence recognition by listeners with cochlear implants. *J Speech Lang Hear Res*, *65*, 3966–3980.

Winn, M. B., Wendt, D., Koelewijn, T., Kuchinsky, S. E. (2018). Best practices and advice for using pupillometry to measure listening effort:

An introduction for those who want to get started. *Trends Hear*, *22*, 2331216518800869.

Wisniewski, M. G., Zakrzewski, A. C., Bell, D. R., Wheeler, M. (2021). EEG power spectral dynamics associated with listening in adverse conditions. *Psychophysiology*, *58*, e13877.

Wingfield, A., Tun, P., Koh, C., Rosen, M. (1999). Regaining lost time: Adult aging and the effect of time restoration on recall of time-compressed speech. *Psychol Aging*, *14*, 380–389.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Chapman and Hall/CRC.

Zekveld, A. A., Kramer, S. E., Rönnberg, J., Rudner, M. (2019). In a concurrent memory and auditory perception task, the pupil dilation response is more sensitive to memory load than to auditory stimulus characteristics. *Ear Hear*, *40*, 272–286.

REFERENCE NOTE

1. Schurman, J. (2021). Effects of age, hearing loss, and cognition on discourse comprehension and speech intelligibility performance. *Doctoral dissertation*. https://doi.org/10.13016/doky-cnjw.