ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Effortful Listening Despite Correct Responses: The Cost of Mental Repair in Sentence Recognition by Listeners With Cochlear Implants

Matthew B. Winn[a] and **Katherine H. Teece**[a]

[a] Department of Speech-Language-Hearing Sciences, University of Minnesota, Twin Cities, Minneapolis

ARTICLE INFO

Article History:
Received November 28, 2021
Revision received April 20, 2022
Accepted June 24, 2022

Editor-in-Chief: Peggy B. Nelson
Editor: Rachael Frush Holt

https://doi.org/10.1044/2022_JSLHR-21-00631

ABSTRACT

Purpose: Speech recognition percent correct scores fail to capture the effort of mentally repairing the perception of speech that was initially misheard. This study measured the effort of listening to stimuli specifically designed to elicit mental repair in adults who use cochlear implants (CIs).
Method: CI listeners heard and repeated sentences in which specific words were distorted or masked by noise but recovered based on later context: a signature of mental repair. Changes in pupil dilation were tracked as an index of effort and time-locked with specific landmarks during perception.
Results: Effort significantly increases when a listener needs to repair a misperceived word, even if the verbal response is ultimately correct. Mental repair of words in a sentence was accompanied by greater prevalence of errors *elsewhere* in the same sentence, suggesting that effort spreads to consume resources across time. The cost of mental repair in CI listeners was essentially the same as that observed in listeners with normal hearing in previous work.
Conclusions: Listening effort as tracked by pupil dilation is better explained by the mental repair and reconstruction of words rather than the appearance of correct or incorrect perception. Linguistic coherence drives effort more heavily than the mere presence of mistakes, highlighting the importance of testing materials that do not constrain coherence by design.

Word and sentence recognition are standard outcome measures in an audiologist's practice. These tasks provide useful data to the patient, clinician, and family that help describe the health of a patient's hearing. However, the speech intelligibility score fails to capture the effort of mentally repairing the perception of speech that was initially misheard. Individuals with hearing difficulty could work backward to mentally reconstruct a sentence in which a word was misheard, leading to successful repetition of a sentence. However, this strategy might also put them at risk for missing the next sentence in continuous conversation. Furthermore, mental repair of speech could have the downside of concealing the struggle that a person undergoes to achieve that understanding because it might appear on the surface that perception was successful. This is an important concept in the field of cochlear implants (CIs) because intelligibility scores are used to determine candidacy, track outcomes, and motivate clinical interventions. To best understand who is in need of specialized audiological care and to better understand the difficulties of listening with hearing loss, it is important to recognize not just the accuracy of repeating speech but also the effort of formulating coherent responses when the input was unclear.

CIs provide a distorted and unclear signal because of various technological and surgical limitations of electrically stimulating the auditory nerve. Arguably, the most glaring limitation is the lack of specific frequency coding because of a limited number of electrode contact sites, and interaction between electrode activation patterns (Wilson & Dorman, 2008). As a result, speech perception

Corresponding to Matthew B. Winn: mwinn@umn.edu. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

can be very difficult, particularly when there is background noise (Gifford & Revit, 2010). However, language knowledge and cognition can play a vital role in successful hearing with a CI. For example, perception of sentences is significantly better than perception of individual words (Gifford et al., 2008) because sentences offer contextual clues and linguistic coherence that CI listeners use extensively (O'Neill et al., 2021; Winn, 2016; Winn & Moore, 2018). A person's ability to exploit contextual clues can be beneficial, but it can also come at a cost of increased effort if the context needs to be used retroactively (Winn & Teece, 2021). This study was designed to track the increase in effort that results from an adult CI user needing to repair a misperceived word in a sentence using later context.

Pupillometry as a Measure of Listening Effort

Pupil dilation has a long history of being associated with changes in cognitive load across a wide variety of tasks (Beatty, 1982; Zekveld et al., 2018). In general, the pupil will dilate more when there is more effort exerted, although there are some caveats to this pattern that are relevant for experimental design (Steinhauer et al., 2022; Winn et al., 2018). As long as the visual environment is controlled to maintain consistent luminance and as long as there is sufficient motivation to exert effort, there is typically a systematic relationship between the difficulty of a task and the increase in pupil dilation in the moments after a stimulus. This method is especially useful for listening tasks, as an auditory stimulus would not interfere with the measurement of pupil size (conversely, visual stimuli can present more complications). In general, pupil dilation can be framed as an index of the need for decision making (Lempert et al. 2015; Satterthwaite et al., 2007) or the recruitment of "extra" resources that force a listener to break from rapid automatic processing (Lemke & Besser, 2016). Pupil dilation reflects the adaptive gain functions driven by the locus coeruleus-norepinephrine system (Aston-Jones & Cohen, 2005) and can be understood as having at least two main components. There is a baseline pupil size that reflects the level of arousal or alertness (McGinley et al., 2015) and the phasic change in pupil size that reflects task-evoked responses (Beatty, 1982). The short phasic changes are used more often as the signatures of listening effort in many publications (Zekveld et al., 2018).

In addition to elevated effort due to decision-making and linguistic processes in cognitive task, there is also a potential impact of stimulus degradation by itself. Systematic degradation of the spectral resolution (clarity) if speech results in corresponding systematic increase in pupil dilation (Miles et al., 2017; Winn et al., 2015), which interacts with tasks involving linguistic comprehension (Winn, 2016). The introduction of noise might also simply raise arousal and therefore increase pupil size, although it is sometimes unclear whether outcome measures are affected by noise level or by intelligibility (Zekveld et al., 2010). These trends, though not perfectly clear, have some implication for the interpretation of effort measures in listeners with CIs because auditory distortion is an inescapable aspect of using the device. However, comparison of pupil dilation measures in CI and non-CI listeners does not show a consistent trend of *overall* increased effort in experiments with constrained listening conditions in quiet (Winn, 2016; Winn & Moore, 2018).

A Focus on Language Processing Rather Than Raw Intelligibility Score

Intelligibility score (repetition accuracy) has been shown to not be a reliable indicator of listening effort in adults in cases of using a CI (Winn & Teece, 2021), listening to degraded speech (Winn et al., 2015), and listening to nonnative accented speech (McLaughlin & Van Engen, 2020). This study design follows up on the study by Winn and Teece (2021), which used sentence stimuli specifically designed to demand retroactive mental repair of distorted words, in which later context provided clues to the identities of those words. In this design, the specific act of mental repair was prospectively built into the study rather than letting it emerge randomly from various difficult listening situations. In addition to regular stimuli that were articulated and heard normally, there were sentences with a mispronounced word that simulated a phonetic misperception; the listener had to correct this word and could, in theory, benefit from phonological similarity of the mispronounced word because only one phoneme changed. Other stimuli had the target word entirely replaced by noise, which was a distortion that was easier to detect because of the acoustic discontinuity, but more challenging to replace the word because there were no perception cues for the word except its duration and intensity, which are not particularly informative. The results clearly showed the effortful cost of repairing a misperception as observed through changes in pupil dilation that were time-locked to the moments of word ambiguity and repair. When a word was mispronounced, there was a brief but significant elevation in effort, which was more substantial and longer lasting when the entire word was replaced by noise. Importantly, these patterns emerged in adults with normal hearing (NH) despite perfect intelligibility scores, suggesting that correct repetition does not protect a listener from the effort of processing and mentally repairing a misperception.

Now, we turn to generalize this study design to adults who have CIs, for whom we expect that this mental activity reflects everyday listening rather than a manufactured

laboratory task. Although a variety of types of listening effort are possible, it is the effort related to language processing and the search for coherence that we focus on here, as it had the greatest impact across a wide variety of error response types in the analysis by Winn and Teece (2021). Consistent with previous reports by Wingfield et al. (1995) and by Potter and Lombardi (1990), errors are commonly found to be semantically coherent with other words in the sentence, even if they were not a good acoustic match (e.g., "My family's Christmas tree has lots of ornaments" repeated as "My family's cake has nuts and almonds"). Additionally, when one word is misperceived, other words elsewhere in the sentence were seen to be forced into alignment with the first misperception to preserve coherence at the expense of preserving acoustic match. This pattern was foreshadowed by analysis by Wingfield et al. (1995) who found that incoherent responses were rare, occurring only 3% of the time. They observed "the addition of words, or the substitution of one word for another, that represented an active reconstruction of the original utterance serving to keep the responses syntactically and semantically coherent." This type of effort reflects active cognitive control (Shenhav et al., 2017) and should emerge more strongly for sentence-length materials when the listener can add something that was not present in the original signal (Rönnberg et al., 2019).

The Invisibility of Mental Repair of Speech

The act of mentally repairing a misperception cannot be ascertained on an audiogram because there is no explicit recognition of the supporting mental processes that have taken place. It would be reasonable to suspect that audiologists or other conversation partners can notice listening effort when talking to someone who is hard of hearing, but there is a need for empirical examination of that ability. Verbal reaction time has been shown to increase in difficult listening situations for children (McCreery & Stelmachowicz, 2013), but it remains unclear whether such reaction time differences are reliably perceptible to an external observer and reliably interpreted as indicating extra effort. Wingfield et al. eloquently summarized the implication for clinical assessment, noting that *"what may appear as a successful reproduction may in fact have been the result of a successful reconstruction*." A rising intonation may suggest that the listener is inviting the talker to clarify incomplete information (Lai, 2010), but the expression and self-awareness of listening effort can be variable across different types of people, who might be more or less willing to describe their difficulties (Kamil et al., 2015). Therefore, there is a need to explore changes in effort within individuals in a systematic prospective experiment that does not rely solely on subjective reporting.

Hypotheses

Speech recognition can appear successful but might have resulted from an effortful process of mental repair of an incomplete perception. Based on combined decades of conversations with audiology patients by the authors and based on the findings of previous studies described above, this study was driven by the following hypotheses.

1. Even when repeating a sentence correctly, listening effort will be higher in cases when the listener had to mentally repair a word using later context.
2. CI listeners will show prolonged elevation of listening effort beyond the end of a sentence, based on previous findings by Winn and Moore (2018) and Winn and Teece (2021).
3. Compared to people with NH, CI listeners will not be as sensitive to the mispronunciations because of their reduced sensitivity to acoustic cues that signal consonant place of articulation (cf. Herman & Pisoni, 2003; McMurray et al., 2019).

Method

Participants

Data are reported here for 17 adults with one or two CIs (4 men, 13 women; age range: 23–81 years; average: 62.5 years). All participants were native speakers of North American English. Demographic information for the included participants is listed in Table 1. Data were collected for three additional participants but were ultimately excluded due to tracking difficulties (details later).

Stimuli

Stimuli included 120 sentences written and recorded by our laboratory (see Winn & Teece, 2021, for a detailed description). Each sentence was designed to have 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. 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 (described by Winn & Teece, 2021). The sentences were divided into four lists of 30, with the average word length of the sentence and average target word position within the sentence equalized across lists. The sentences had an average of nine words, and the target word occurred, on average, at word position 3.35. Because of the goal to systematically mask words in isolation, it was important that the sentences are highly intelligible to minimize the tendency to make

Table 1. Demographics of CI participants.

Listener	Sex	Age	Device type	Ear(s)	Etiology	CI Exp. (y)
C115	F	81	Cochlear	Bilateral	Idiopathic SNHL	7.5
C118	F	30	Cochlear	Bilateral	Idiopathic SNHL	8
C119	F	23	Cochlear	Bilateral	ANSD	17.5
C126	F	72	Med-El	Bilateral	Idiopathic SNHL	5.5
C130	M	66	Med-El	Right	Genetic SNHL	1
C131	F	70	Cochlear	Right	Chronic middle ear disease	5.5
C134	F	63	Cochlear	Bilateral	Idiopathic SNHL	6
C138	F	60	Advanced Bionics	Bilateral	Idiopathic SNHL	28
C139	F	61	Advanced Bionics	Bilateral	Genetic SNHL	7.5
C141	F	73	Advanced Bionics	Right	Genetic SNHL	7
C143	F	64	Cochlear	Bilateral	Bacterial labyrinthitis	3
C144	F	62	Cochlear	Bilateral	Measles	16
C145	M	54	Cochlear	Bilateral	Meniere's disease	6
C146	F	67	Cochlear	Bilateral	Idiopathic SNHL	7
C147	F	71	Cochlear	Right	Barotrauma	1
C148	M	70	Cochlear	Left	Otosclerosis	2
C156	M	76	Cochlear	Right	Chronic middle ear disease	1

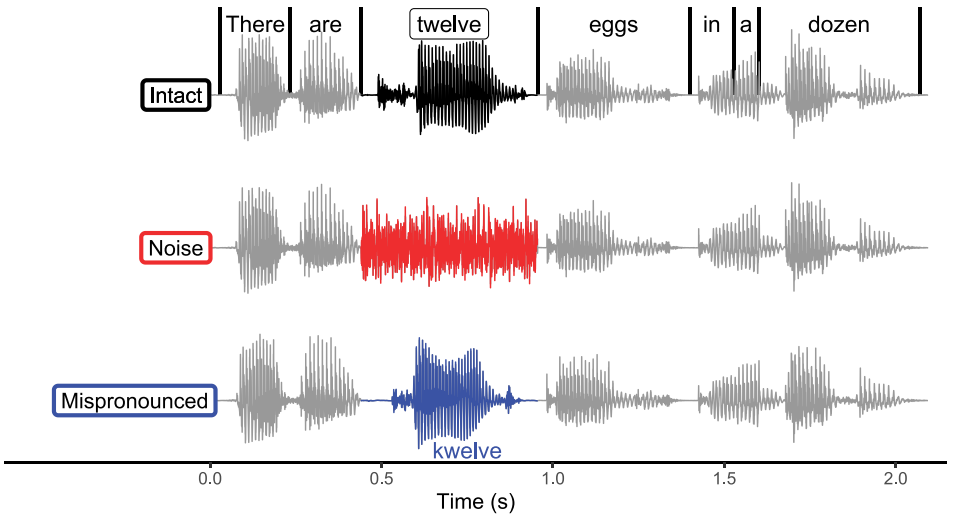
Note. CI = cochlear implant; Exp. = experience; F = female; SNHL = sensorineural hearing loss; ANSD = auditory neuropathy spectrum disorder; M = male.

mistakes on words other than the manipulated target word. The sentences were spoken by an audiologist (the first author) with explicit effort to facilitate clear understanding.

Stimulus Variations

There were three versions of each sentence, as illustrated in Figure 1. The "intact" version was the full utterance with all words spoken naturally. There were two versions that distorted the target word, which forced the listener to engage in some mental repair. In the "Noise" condition, the target word was replaced with noise matched in duration and intensity, whose frequency spectrum matched the long-term spectrum of the entire stimulus corpus. The second type of distortion was an intentional mispronunciation of the first consonant in the target word. This mispronunciation nearly always was a change in the place of articulation of the consonant, which is the feature most often misperceived by listeners with hearing loss (Dubno et al., 1982), including those who wear CIs (Munson et al., 2003; Rødvik et al., 2019). Most of the mispronunciations resulted in nonwords. The mispronunciations were spoken with the same prosody by the same talker and spliced onto the intact form of the sentence starting from the end of the target word. The goal of this splicing was to ensure that the audio content

Figure 1. Types of stimuli used in the sentence recognition experiment, including an "intact" form (top) where all the words were unaltered, a "noise" form (middle) where a single target word was replaced with noise of equal duration and intensity, and a "mispronounced" form (bottom) where a phoneme in the target word was mispronounced.



following the target word, which served to disambiguate the target word itself, was exactly the same in all versions of the stimuli.

Procedure

Prior to data collection, this experiment was reviewed and approved by the University of Minnesota institutional review board. Written consent was obtained for all participants. Each participant completed a sentence-repetition task with a total of 120 stimuli (40 sentences each for intact, noise-masked, and mispronounced trials). These stimuli were divided into four blocks of 30 sentences each. 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 (intact, noise-masked, or mispronounced) for each item was rotated for each listener, except for the first trial in each list.

During the experiment, listeners sat in a chair with their forehead position stabilized by the upper bar of a chinrest whose base was sufficiently lowered to allow comfortable jaw movement for speaking. They visually fixated on a red cross in the middle of a medium-dark gray background on a computer screen that was 50 cm away. Each trial was initiated by the experimenter, and the participant heard a beep marking the onset of the trial. There was 2 s of silence, and then 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. They were instructed to repeat back what they thought was spoken, filling in missing or distorted words when necessary. The participants' verbal responses were scored on paper and also audio recorded for later inspection. Incorrect responses were saved for further analysis of error patterns. 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. Lighting in the testing room was kept constant.

Analysis

Intelligibility

Intelligibility for each word in the sentence was scored in real time by an experimenter, and the responses were audio recorded for later inspection. For stimuli whose target was replaced by noise, any word that was not semantically coherent with the stimulus was counted as an error, as well as any errors elsewhere in the sentence. If the participant's guess at the word replaced by noise was not the "intact" version of the word but still made sense (e.g., "The worker *used* the ladder to get to the roof" instead of "The worker *climbed* the ladder to get to the roof"), it was counted as correct. We also tracked whether participant responses were linguistically coherent and the presence of multiple errors within trials.

Pupillometry Data Preprocessing

Pupil data were processed in the style described by Winn et al. (2018) and Winn and Teece (2021). 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 ms and forward by 120 ms to account for the partial occlusion of the pupil by the eyelids during blinks. 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 ms before stimulus onset to 500 ms 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 s past the onset of the stimulus. CI listeners on average had fewer trials discarded due to missing data (1.7%) and less variation among individuals (SD = of 13%) compared to NH listeners from the previous study (average of 3.1% trials discarded with SD = of 17%). Other outliers/contaminations were automatically detected through an algorithm that accumulated multiple "flags," such as high-intensity hippus activity during baseline, baselines that had extraordinary deviation from both the previous and the next baseline, significant slope of change in pupil size during the baseline, or a significant negative swing in proportional dilation immediately after the stimulus onset. Three or more flags resulted in a trial being dropped. If a participant had fewer than 13 trials remaining in any condition following outlier detection, that participant's entire data set was dropped; two NH listeners and one CI listener were excluded for this reason. Among the 34 participants remaining in the data set, 5.1% of trials were discarded, with more for NH listeners in the previous study (6.8% with SD = 5.8%) compared to CI listeners (3.2% with SD = 3.2%), with the greatest number of rejections for NH listeners in the Intact condition, suggesting that the easiest trials produce the most problematic data.

Pupillometry Data Analysis

Filtered data that were summarized for each individual in each stimulus condition were estimated using a second order (quadratic) polynomial model (see the studies of Mirman, 2014; Winn et al., 2015). An alternate model using individual trial-level data was attempted but ultimately abandoned because the requisite computing power and model complexity was not justifiable by the data. Consistent with the previous analyses in similar studies,

there were two windows of analysis, intended to treat audition and linguistic processing as two separate processes rather than a singular process. Window 1 spanned from −1.5 to 0.7 s relative to sentence offset, which corresponded to the *listening* phase of each trial. On average, the first analysis window began about 800 ms after the onset of the sentence, which captures the onset of the response while accounting for the roughly 700-ms delay in dilation upon response to the sound. Window 2 spanned from 0.7 to 2.2 s relative to sentence offset, reflecting the *response preparation* phase of the trial. These windows arguably correspond to auditory encoding versus poststimulus linguistic resolution, and they have been separately analyzed in numerous previous studies that find distinctly different effects in each window (Bianchi et al., 2019; Francis et al., 2018; Piquado et al., 2010; Wendt et al., 2016; Winn, 2016; Winn & Moore, 2018; Winn & Teece, 2020, 2021; Winn et al., 2015). Peelle and Van Engen (2021) comment on the value of close inspection of time windows in this style of analysis when drawing conclusions from the analysis.

Within each analysis window, there were fixed effects of stimulus type and time. There was a maximal subject-level random-effects structure, meaning for each fixed effect, there was a corresponding random effect declared per listener to account for dependence between repeated measures and between samples of the same measure over time. The prevailing model formula took the following form for each analysis window:

$$\begin{aligned}&\text{lmerTest :: lmer(pupil}\sim\text{poly1} + \text{poly2} + \text{Type} + && \#\textit{Main effects}\\ &\quad \text{poly1} : \text{Type} + \text{poly2} : \text{Type} + && \#\textit{two-way interactions}\\ &\quad (1 + \text{poly1} + \text{poly2} + \text{Type} + && \#\textit{main random effects}\\ &\quad \text{poly1} : \text{Type} + \text{poly2} : \text{Type}|\text{Listener}), && \#\textit{random interactions}\\ &\text{data} = \text{Data\_window})\end{aligned} \quad (1)$$

... where poly1 and poly2 are orthogonal polynomial transformations of time relative to stimulus offset, and Type is stimulus type, with "mispronounced" as the default configuration. "Data_window" is the subset of data from within either the first or the second time window. Polynomial time transformations were used so that three separate properties of the dilation curve—absolute level, slope, and curvature—could be assessed separately. The orthogonal nature of this transformation meant that the polynomials were phase-shifted so that the properties were not correlated with one another (as opposed to typical phase-aligned polynomials, where the growth of a number x would be correlated with the growth of x^2). The linear term (poly1) corresponds to the growth rate of pupil size, and the quadratic effect (poly2) corresponds to the curvature (or deceleration) of growth in pupil size. All data and code to run analyses and plotting can be found on an Open Science Foundation (OSF) site located at: https://osf.io/ctnrj/.

Results

Intelligibility

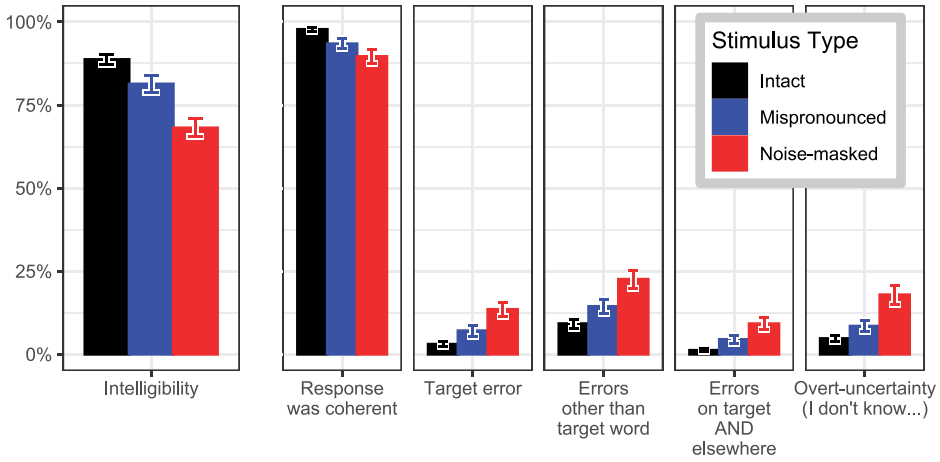
Intelligibility scores were high for all sentences, with performance at 80% overall, confirming that performance did not dip into the range where motivation and effort are in doubt (cf. Wendt et al., 2018). By stimulus type, intelligibility was 89% correct for intact sentences, 82% for mispronounced sentences, and 68% for noise-masked sentences. See Figure 2 below for those patterns, as well as the percentage of time that sentences were coherent, when errors were made on words other than the target, and trials when the participant overtly indicated that they knew they did not process the sentence accurately. Across the 40 trials, the participant overtly indicated uncertainty (replacing a word with "something" or saying "umm, I didn't catch that word") on average 1.8 times for the intact stimuli, 3.1 times for the stimuli with mispronounced words, and 6.3 times for the stimuli with noise-masked words. Analysis of variance revealed an effect of stimulus type ($F = 33.4$, $p < .001$) on the tendency to signal uncertainty, and follow-up comparisons revealed each stimulus type to be statistically different from each of the others. However, participants varied widely in their tendency (or *willingness*) to express uncertainty, with overt-uncertainty-signaling counts ranging from zero (C126) to 25 out of 120 (for C146).

There was a trend for older listeners to do more poorly on the noise-masked target stimuli in terms of overall intelligibility, but this trend was not statistically detectable (interaction of age and stimulus type: $t(30) = -1.69$, $p = .10$) using a linear model estimating intelligibility as a function of age, stimulus type, and the interaction between those two terms, with a random intercept for each participant. The trend that was attributable to increased prevalence of errors on words all across the sentence not just the target words. There was no effect of participant age on any of the other intelligibility markers (target-specific errors and coherence of responses). Because the participant pool did not uniformly sample the age range and because the statistical model could not accommodate a full fixed-effect structure (note that the degrees of freedom = 30 rather than 16), we refrain drawing strong conclusions about these particular trends.

Verbal Reaction Times

Verbal reaction times are displayed in Figure 3. Reaction times below the third or above the 97th percentile were trimmed out of the analysis (130 data points out of 2,159 were excluded). A linear mixed-effects model describing these data included main effects of stimulus type and overt-uncertainty, but an interaction between these effects was shown to not improve the model (increase in Akaike information criterion, nonsignificant Chi-square statistic of 0.45). The only random

Figure 2. Intelligibility patterns for the sentence recognition task by stimulus type. The first and second panels indicate the percent correct and coherence of responses. The third panel displays errors on the target word. The fourth panel displays frequency of errors made not on the target word but on another word in the sentence. The fifth panel displays the occurrence of multiple errors in the sentence. Finally, the sixth panel reveals how many times a participant stated out loud that they could not repeat all or part of the sentence definitively.

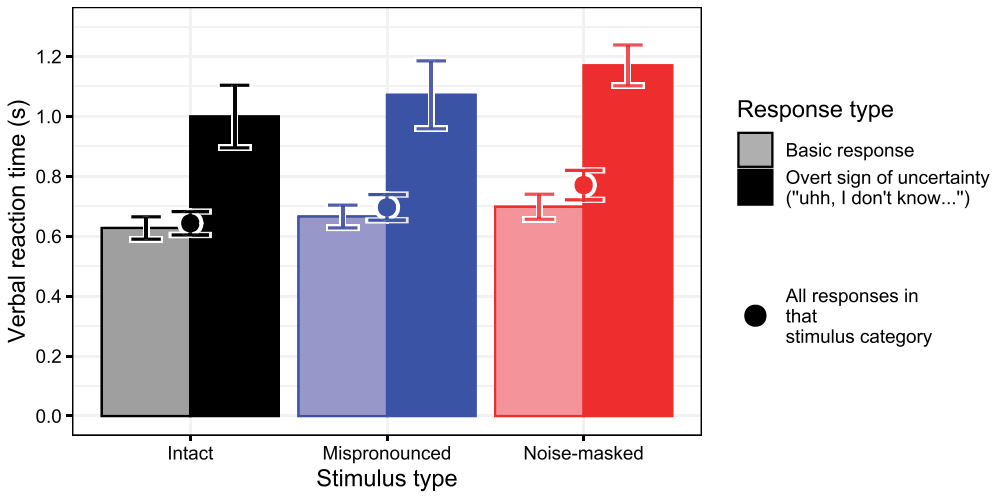


effect was a random intercept per listener. There was a significant effect of stimulus type, with mispronounced target words eliciting longer reaction times that were roughly 39 ms longer than those for intact stimuli ($t = 2.51$, $p = .012$), and noise-masked stimuli eliciting reaction times that were 80 ms longer than those for intact stimuli ($t = 5.09$, $p < .001$). Overt uncertainty had the largest effect, increasing the reaction time by an average of 376 ms ($t = 17.13$, $p < .001$).

Pupillometry

Figure 4 displays the changes in pupil dilation elicited by the different types of stimuli, and it provides comparison to the data collected in listeners with NH in our earlier study along with a comparison to data for only correct trials (in dashed lines). The overview of the results is that, when compared to fully intact sentences, sentences with mispronounced words elicited slightly larger pupil dilation in the moments after the sentence was complete. Sentences with noise-masked target words elicited substantially greater and earlier pupil dilation, which persisted through the retention interval between sentence and response. In contrast, the NH listeners responded to the mispronounced stimuli with a larger but briefer increase in pupil dilation that quickly converges back to the same pattern as that for the Intact stimuli. Average responses to

Figure 3. Verbal reaction times in seconds by stimulus type and degree of certainty. The left-sided lighter bars indicate the times for sentences repeated in a confident manner, and the right-sided darker bars indicate the times for sentences spoken with an overt sign of uncertainty.



the sentences with masked words were nearly identical across groups, but had slower recovery toward baseline for the CI group; this and other patterns were statistically analyzed in the sections that follow.

The model for pupil responses during Window 1 (−2 to 0.7 s relative to sentence offset) used the Mispronounced stimulus type as the default, to compare the Intact condition downward and the Noise condition upward. In general, the responses to the Mispronounced stimuli were more like those to the Intact stimuli than to the Noise stimuli. The interaction of stimulus type with the intercept was shown to be statistically greater for the Noise condition, $t(16.6) = 6.3$, $p < .001$, and smaller for Intact stimuli, $t(15.9) = -2.35$, $p = .03$. The slope of pupil dilation across time was statistically steeper when the stimulus type was Noise, $t(16.2) = 4.13$, $p < .001$, and statistically lower when the stimulus was Intact, $t(15.7) = -5.09$, $p < .001$. The quadratic term, reflecting the deceleration of the growth of pupil dilation, was statistically detectable for the Mispronounced condition, $t(16.4) = -2.74$, $p = .014$. The quadratic term for the Intact stimuli was not different from that of the mispronounced stimuli, $t(16.1) = 0.28$, $p = .78$, but was larger (more negative) for the noise condition, $t(16) = -3.7$, $p = .002$, reflecting greater deceleration accompanying the more rapid dilation for the Noise stimuli.

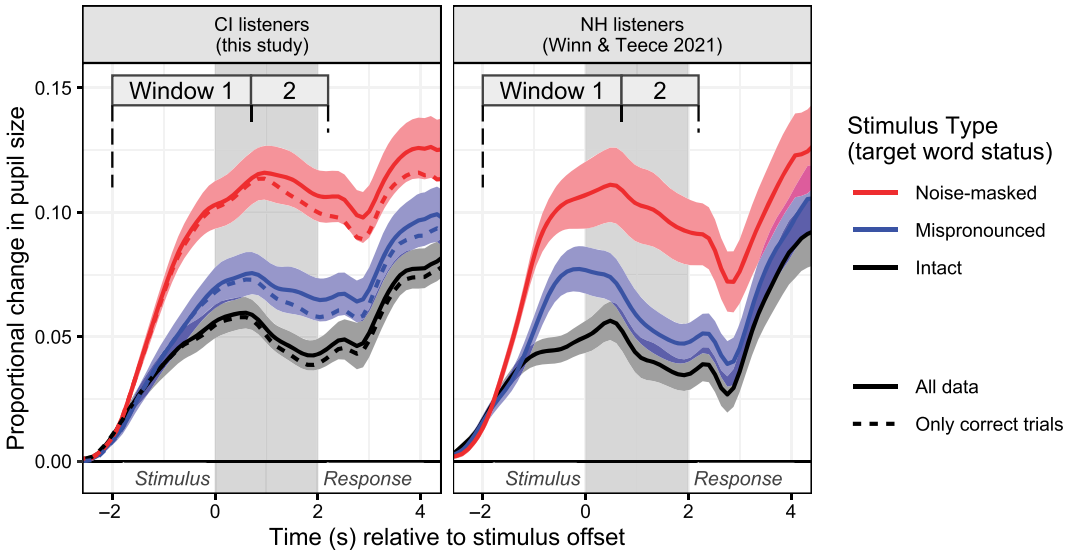
Tracking the Effort of Repairing Words

Window 2 used the same prediction terms as those for Window 1 but in a different chunk of time (0.7–2.2 s relative to sentence offset). This model estimated the flat/downward slope away from peak pupil dilation up to the point of the visual response prompt. Most of the differences between conditions were captured in the intercept terms, as the Intact condition elicited statistically smaller dilation, $t(15.7) = -2.95$, $p = .009$, and the Noise condition yielded statistically larger dilation, $t(15.9) = 6.18$, $p < .001$, compared to the default Mispronounced condition. The slope of the dilation curve for mispronounced stimuli was marginally negative, $\beta = -0.012$, $t(16.9) = -2.21$, $p = .04$, and the slopes for the other conditions were not statistically different. This stands in contrast to results reported earlier for listeners who have NH who show a steeper downward slope (recovery to baseline) for both the intact and mispronounced stimuli compared to the noise stimuli (Winn & Teece, 2021). There were no statistically detectable quadratic effects during Window 2. Further analysis into the mispronounced stimuli revealed no differences in trials when the mispronounced word became a nonword versus another real word (plot and analysis code available in the supplemental materials on OSF).

Modeling Responses When Answers Are Correct

Previous studies have shown that pupil responses tend to be smaller in sentence-recognition tasks when verbal responses are correct (McHaney et al., 2021; Winn, 2016; Winn et al., 2015; Zhang et al., 2021). A follow-up analysis was done for the current data to compare the results from the full data set versus only trials with correct responses, with *data set* interacting with each of the terms in the original statistical model. There were no statistically

Figure 4. Proportional changes in pupil dilation elicited by different types of stimuli. The left panel shows data in this study obtained from listeners with cochlear implants (CIs), with a subset of data in dashed lines from correct trials. The right panel shows data from the same experimental design obtained in listeners with normal hearing (NH) from a previous study, for comparison.



detectable effects of response correctness in the first window analysis. In the second window, there was a significantly lower overall dilation, $t(13.9) = -3.41$, $p = .004$, and also steeper recovery back to baseline (negative slope interaction) for the mispronounced condition, $t(21.7) = -2.79$, $p = .011$, for correct trials, an interaction that was not different for the other two conditions ($p = .13$ and .90 for intact and noise, respectively). This pattern suggests that correct responses led to more recovery back toward baseline that was essentially independent of stimulus type.

Comparing CI Data to NH Data

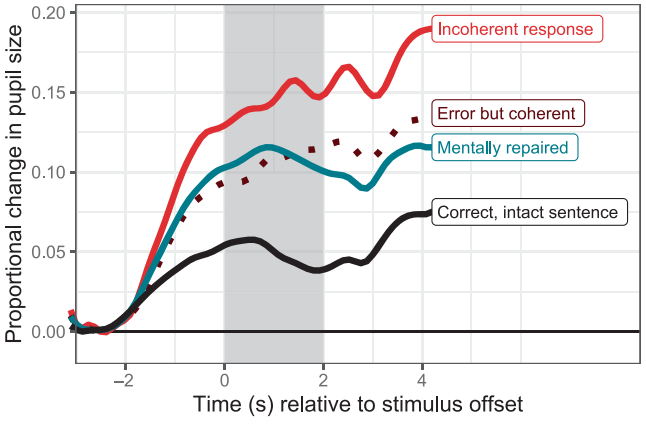
Based on the summary data shown in Figure 4, the biggest difference between the CI group and the NH group was the response to the mispronounced stimuli. NH listeners showed momentary increase in dilation that quickly drops back down to converge back with the responses for the Intact stimuli, whereas the CI listeners show a smaller increase in dilation to mispronounced stimuli, but that extra dilation sustains for a longer period of time. These impressions were confirmed by the statistical analysis, which revealed greater curvature of pupil dilation in NH listeners for this stimulus type as reflected by the quadratic term, $t(31.8) = 2.13$, $p = .04$. The difference between quadratic effect magnitude across the Intact and Mispronounced conditions was stronger for NH listeners as well, $t(29) = 2.72$, $p = .011$, consistent with the overall stronger curvature in the data for the mispronounced type of stimuli. These were the only terms that interacted significantly across hearing groups for Window 1. During Window 2, the only statistical interactions with hearing group were for the quadratic term for the mispronounced stimuli, indicating more curvature for NH listeners, $t(31.22) = 2.17$, $p = .04$, and a reduced quadratic term for the Intact stimuli for NH listeners, $t(33.4) = -2.56$, $p = .015$. Otherwise, the pupil size changes did not interact with hearing.

Figure 4 suggests that the lowered pupil dilation for CI listeners in Window 2 for correct trials rendered their data visually similar to the data for NH listeners. Nevertheless, the same pattern of effects and noneffects of hearing generally persisted even in a follow-up model that included only correct trials. The only potential difference was that the increase in curvature in dilation in response to mispronounced stimuli in NH listeners fell fellow a conventional significance criterion, with p value changing from .04 to .07. There was no difference between NH and CI listener responses when restricting analysis only to the trials with noise-masked target words (regardless of response correctness). This is consistent with the noise burst being equally easy to detect for both groups, as opposed to the mispronounced words, which might not even be detected by the CI listeners in some trials.

The Influence of Response Demands and Linguistic Coherence

Figure 5 displays the average pupil responses for trials split by factors relating to the mechanism of sentence processing rather than by trial type in an attempt to model the effort of various paths to understanding the speech. Correct refers to correct responses to Intact sentences; these responses elicited the smallest pupil dilation. "Mentally Repaired" refers to sentences with noise-masked target words when the participant's response included a sensible word in place of the noise. "Error but coherent" refers to either (a) intact or mispronounced sentences with any error that was still sensible in context or (b) sentences with noise-masked target words with a sensible error anywhere other than the target word. An example of a sensible/coherent error is the stimulus, "The referee blew the whistle to call the foul," repeated as, "The referee decided to call the foul," or the stimulus, "I got a flat tire when driving on a bumpy road," repeated as, "I got a flat tire when driving on an empty road." Finally, "Incoherent response" refers to responses that did not have full syntactic structure or were not semantically sensible (e.g., the sentence, "the child felt a bee sting her on the arm," repeated as, "The child felt to be still her," or the sentence, "I only want one serving, because I'm not that hungry," repeated as, "The room is assertive as I met my hungry"). As shown in a previous study with CI listeners in a different task, incoherent answers elicit the largest pupil dilation on average, although the statistical regularity of this pattern is difficult to discern, because these responses are smaller and more variable in number across the participants. Coherent errors elicited nearly the same dilation as corrections of missing words, implying that it was the reconstruction—not the

Figure 5. Proportional changes in pupil dilation elicited by stimuli hypothesized to demand different kinds of mental processing. Each line contains a variable number of trials obtained the participants, and it is meant only as an introductory glimpse at patterns that were either planned (Correct/mentally corrected) or unplanned (Error but coherent/Incoherent).



ultimate match to target stimulus—that drove the pupil response. When no sensible word could be produced during that reconstruction, dilation was further elevated. Keeping in mind the limitation of full analysis, these patterns loosely suggest that the act of mental search for coherence—not the status of the response as an error—is what drives listening effort during sentence recognition.

General Discussion

This study reinforces the main conclusion of Winn and Teece (2021) and extends it to adults who use CIs. The most important finding in this study is that there is a significant increase in listening effort in CI listeners if part of a sentence was mentally repaired, *even if the verbal response is ultimately correct* (validating Hypothesis #1). The general implication is that listening effort should not be understood as a simple product of the intelligibility mistakes because effort can increase even when no mistakes are clearly present in a verbal response. Another notable finding is that the pupil responses for both intact and noise-masked trials were remarkably similar for the listeners with NH (data from the previous study) and the CI listeners (in this study). These patterns suggest that CI listeners might not experience a *persistent* increase in effort when listening, but instead, they might need to exert effort on a short-term basis (for mental repair) more frequently.

Postpeak pupil dilation offset slopes were shallower for CI listeners (see Figure 4), validating Hypothesis #2. This apparent prolonged effort is consistent with our earlier studies (Winn, 2016; Winn & Moore, 2018) and studies from other laboratories (Farris-Trimble et al., 2014; McMurray et al., 2017) and consistent with anecdotal reports of prolonged uncertainty described by a majority of our CI participants. Indeed, the persistence of poststimulus pupil dilation has been observed across a wide range of tasks relating to the uncertainty of decision-making (Lempert et al., 2015; Satterthwaite et al., 2007). The current results demonstrate a very large difference in the time course of resolving a misproduced phoneme. NH listeners showed a short burst of dilation with fast recovery, within less than a second after stimulus offset. Conversely, CI listeners showed a weaker response to those mispronunciations (validating hypothesis #3) but showed slower recovery after detection, with dilation curves still unresolved almost 3 s after stimulus offset. The likely reason for the reduced effect in the CI listeners is that some of the mispronunciations likely went unnoticed because of the auditory distortion, consistent with a single-subject report by Herman and Pisoni (2003) and consistent with lexical-decision data by McMurray et al. (2019). In fact, the delay in the elevated pupil dilation for CI listeners in response to mispronounced words was almost identical to the 900 ms identified by McMurray et al. (2019). But the increased effort, when it occurred, was likely prolonged because of the listener's inability to judge whether it was a mispronunciation versus their own misperception. Conversely, the NH listeners should be more likely to attribute the distortion to the talker rather than their own hearing.

In this study, the linguistic coherence of participant responses had a larger effect on effort than any of the prospective stimulus manipulations. It is worth reflecting on the implications of this finding for stimulus selection in other experiments on listening effort. In cases where incoherence would be precluded by the design of the study (e.g., when stimuli are digits, single words, or phonemes), a substantial source of effort would be overlooked by design. Consistent with pupil dilations reflecting decision-making processes (Sattherthwaite et al., 2007), we would expect weaker responses in situations when stimuli do not demand actionable decisions of in cases when misperceptions do not evoke additional contemplation because they are already coherent.

Misperceptions Are Contagious

Although it was not surprising that listeners with CIs made more errors overall, a novel result was that the need to mentally repair a word resulted in a greater prevalence of errors *elsewhere* in a sentence (see Figure 2). There are at least two explanations of this result according to previous studies with adults. The first is that errors tend to be coherent with other words in a sentence; hence, an additional mistake might be generated in order to be coherent with an earlier mistake (i.e., a *type-5 response*, cf. Wingfield et al., 1995; Winn & Teece, 2021). Although not a statistically strong effect, there was a tendency for CI listeners to make more mistakes on later words following masked targets. This is consistent with results by Wingfield et al. (1994) who found retroactive context to be less effective for older adults, and a good number of participants in this study were older than those in the NH comparison group. The second explanation is that the effort needed to repair one misperceived word—even if that process is successful—drains resources that could have otherwise been used to successfully attend to other words in the sentence. This pattern likely reflects the experience reported often by individuals who have hearing difficulty, which is that the contemplation of a single word can last long enough to interfere with the ability to follow conversation smoothly. Consistent with this idea, Pisoni et al. (2018) observed that adult CI listeners would show deficits on later test items after encoding earlier test items and also that retrieval of items from memory interfered with recall of later items (a phenomenon coined "retrieval-induced forgetting," cf. Anderson et al., 1994). Furthermore, it could explain why reducing listening effort can be

advantageous in situations when continued listening is expected (Winn & Moore, 2018).

An additional implication of the "spreading" of errors in the mentally repaired sentences is that intelligibility for an individual word in a sentence does not necessarily reflect auditory encoding of that individual word. Consider that a single-phoneme substitution would often result in complete incoherence of a word in context, yet Wingfield et al. (1995) observed such errors just 3% of the time in sentence contexts. The tangible impact of this is that experimenters should exercise caution when interpreting an error within a sentence as reflecting errors on the specific phonemes in that word because the mistakes could be driven by semantic coherence more often than by acoustic similarity (Potter & Lombardi, 1990).

Effort as a Momentary State Rather Than as an Ongoing Trait

Listening effort is often framed as if it is a *trait* of a person or a task condition, like the effort associated with a specific background noise level, a spatial configuration of noise, a type of noise, the degree of signal degradation, or degree of hearing loss. The clarity and cohesiveness of the data patterns in this study reinforce the idea that it is worthwhile to understand effort as an event or a momentary *state* rather than just a continuous trait of a person or a situation. Effort will emerge not just from a circumstance but from specific moments that require deliberate use of cognitive resources. This study was designed to control when the listener needed to engage in cognitive repair, rather than providing the situation that would likely (but not definitely) lead to the mistake. Conversely, in many other studies that use a variety of speech degradations (e.g., masking noise, vocoding, and reverberation), it is not always possible to know if specific misperceptions occurred or when they occurred. For example, ongoing masking noise might elicit an error near the beginning of one sentence but near the end of another sentence, and those errors might be hidden from the experimenter if the listener mentally repairs them before verbally responding. Aligning physiological responses to those specific misperceptions becomes unfeasible in such a situation, and the local signatures of effect become spread thin and lost through time-series averaging. There are only a small number of studies that specifically disentangle effort as a general disposition versus effort selectively applied in specific moments (cf. McLaughlin et al., 2018), but this is an area of potential rich exploration and insight.

The complex nature of the stimuli used in this study complicates any comparison to analyses of previous studies that expressed results in terms of peak dilations and peak latencies. The main problem was that individuals have different morphologies of pupil dilation, which interacted with the different morphologies elicited by the stimulus types. In some cases, dilation grew monotonically from the auditory sensation all the way through to the verbal response, resulting in no clear "peak" apart from whatever value was obtained at the end of an analysis window. For the mispronounced stimuli, many listeners demonstrated a localized early bump in dilation, but it was not consistently the *peak* dilation. As a result, peaks for the mispronounced stimuli were not comparable to the peaks for other stimuli for which the dilation grew monotonically. Peak dilations and latencies could be subject to further analysis using the data openly available on this project's OSF website, with the prevailing observation that such extracted features should be used with understanding of whether they reflect something real and consistent in the underlying data.

Limitations of the Current Measurements

The most important comparison in the study is between effort for the Intact sentences versus sentences that demanded mental repair via masking of a target word. However, it is not possible to rule out mental repairs for the Intact sentences because listening with a CI typically leads to some mistakes in perception even in the absence of any stimulus distortions. Even the use of slower and more careful articulation for the sentences did not prevent misperceptions, as can be seen from the intelligibility data. The comparison in the current analysis therefore rests on the assumption that the number of Intact stimuli requiring mental repairs would be infrequent enough that the data would reflect a true difference from the stimuli with masked target words, which required mental repair every single time. The patterns in the data are clear enough to conclude that the current stimuli can bring out the hypothesized differences in effort, but the true effect might be larger than expected, given the tendency for occasional perceptual errors that were not planned in the design.

Although this study was designed to specifically track a form of effort that has been mostly elusive in previous work, there are numerous other domains that contribute to listening effort. Hughes et al. (2021) provide a taxonomy of numerous factors other than pure auditory encoding that should be considered when designing and interpreting studies of speech communication among people who are hard of hearing. Even among listeners with typical hearing, other relevant factors include background knowledge, anticipating others' intentions, sorting through similar-sounding options, handling mistakes, and the effort of deciding whether perception is worth the effort. Effort can arise from the cost of task-switching (McLaughlin et al., 2019), test anxiety and motivation to perform (Jones et al., 2015), awareness of mistakes, or physiological response to noise (Kim et al., 2021; Love et al., 2021), among other factors.

Effort can also be modified by familiarity with topics and similarity of stimuli in a short span of time (Konopka & Kuchinsky, 2015), which was not controlled in this study. Pupil-indexed effort is not the only, or even the most important, kind of effort. Alhanbali et al. (2017, 2019) and Strand et al. (2018) demonstrated a lack of intercorrelation between effort measures, and thus, these measurements might not map cleanly to a person's lived experience, even if the measures themselves are reliable across testing sessions (Alhanbali et al., 2019).

Another limitation of this study is that the responses to the mispronounced stimuli are potentially not fairly comparable across the NH and CI groups. For the NH listeners, the mispronunciations would be easily detectable and dismissed as fault of the talker rather than as a misperception. Conversely for the CI listeners, the mispronounced words—if they were noticed at all—might have caused a momentary state of uncertainty as to whether the errant perception was the fault of the talker or auditory limitations of the listener. Anecdotally, both of these situations arose; a small number of CI participants reported never hearing a misperception, whereas one highly successful CI participant reported overly fixating on the misperceptions while contemplating the very notion of the possibility of an auditory mistake.

The New Open Questions

Three new fundamental questions about listening effort emerge from the current results. First, although this article focuses on the mental effort of parsing and assembling coherence between words, degradation by itself (e.g., the presence of distortions or background noise) might incur at least some elevated physiological response independent of task performance and language processing (cf. Francis & Love, 2020; Francis et al., 2016). We do not know the relative contributions of each of these kinds of effort, or whether they interact. A second fundamental question is whether this study reflects a realistic timescale of mental correction in everyday listening by individuals who use CIs. Each of the sentences in this study was independent; hence, the act of mental repair was rather immediate. In regular conversation, a listener might instead hold off on committing to a perception or a repair until later sentences are heard. Some of our participants informally described a process of perceiving "chunks" of sentences at a time, possibly in order to guard against committing to the wrong path. Nieuwland (2021) highlights the complexity and pitfalls of speculating about how predictions could be suppressed or modified by immediate experience.

A third fundamental question is the extent of individual differences in *willingness* to engage in effortful listening. We expect that listeners in everyday settings can downregulate the mental cost of constructing meaning in sentences by simply choosing to disengage. Rather than being a sign of inattentiveness, this behavior by people with hearing difficulty would be strategic and economical so that the individual can preserve necessary mental resources for more important upcoming situations. For this study and nearly all other studies in the literature, there are low stakes for providing false responses; hence, the real-life anxiety of misperception could be drastically underestimated in the laboratory. By comparison, the stakes (and the accompanying anxiety) could be higher for questions about the listener or conversation that would have impact on work, friendship, or family.

Conclusions

Even if a person with a CI appears to correctly repeat a sentence, there is a significant increase in listening effort if one of the words was misperceived and then mentally repaired before the verbal response. Furthermore, that effort lingers for a longer amount of time in CI listeners compared to effort elicited in people with NH. This means that, in standard clinical and laboratory assessments of CI users, the effort of speech perception can be underestimated because effortful mentally corrected responses are counted the same as effortless correct responses. Effort is further increased when the repair strategies do not work to successfully produce a coherent response. Therefore, stimuli that are constrained to offer no chance to resolve incoherence will not likely elicit the effort that dominates perception of regular open-ended sentences. Counseling ought to raise awareness of effort and fatigue so that patients can take an active role in budgeting their effort and in making intentional decisions about whether and how they want to signal it to conversation partners and health care professionals.

Data Availability Statement

All data and code to run analyses and plotting for this study can be found on an Open Science Foundation (OSF) site located at: https://osf.io/ctnrj/.

Acknowledgments

This research was supported by NIH Grant R01 DC017114 (Matthew Winn). The experiment design was assisted by our late colleague Akira Omaki. Data collection was assisted by Steven Gianakas, Maria Paula Rodriguez, Siuho Gong, Hannah Matthys, Lindsay Williams, Emily Hugo, and Justin Fleming. Valuable input to this project was given by our laboratory participants, as well as by Allison Johnson, Peggy Nelson, and Benjamin Munson. All data and code to run analyses and plotting

can be found at https://osf.io/ctnrj/. The University of Minnesota stands on Miní Sóta Makhóčhe, the homelands of the Dakhóta Oyáte.

References

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing, 38*(1), e39–e48. https://doi.org/10.1097/AUD.0000000000000361

Alhanbali, S., Dawes, P., Millman, R., & Munro, K. (2019). Measures of listening effort are multidimensional. *Ear and Hearing, 40*(5), 1084–1097. https://doi.org/10.1097/AUD.0000000000000697

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1063–1087. https://doi.org/10.1037/0278-7393.20.5.1063

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292. https://doi.org/10.1037/0033-2909.91.2.276

Bianchi, F., Wendt, D., Wassard, C., Maas, P., Lunner, T., Rosenbom, T., & Holmberg, M. (2019). Benefit of higher maximum force output on listening effort in bone-anchored hearing system users: A pupillometry study. *Ear and Hearing, 40*(5), 1220–1232. https://doi.org/10.1097/AUD.0000000000000699

Dubno, J. R., Dirks, D. D., & Langhofer, L. R. (1982). Evaluation of hearing-impaired listeners using a Nonsense-Syllable Test. II. Syllable recognition and consonant confusion patterns. *Journal of Speech and Hearing Research, 25*(1), 141–148. https://doi.org/10.1044/jshr.2501.141

Farris-Trimble, A., McMurray, B., Cigrand, N., & Tomblin, J. B. (2014). The process of spoken word recognition in the face of signal degradation. *Human Perception and Performance, 40*(1), 308–327. https://doi.org/10.1037/a0034353

Francis, A. L., & Love, J. (2020). Listening effort: Are we measuring cognition or affect, or both? *WIREs Cognitive Science, 11*(1), e1514. https://doi.org/10.1002/wcs.1514

Francis, A. L., MacPherson, M. K., Chandrasekaran, B., & Alvar, A. M. (2016). Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Frontiers in Psychology, 7,* 263. https://doi.org/10.3389/fpsyg.2016.00263

Francis, A. L., Tigchelaar, L. J., Zhang, R., & Zekveld, A. A. (2018). Effects of second language proficiency and linguistic uncertainty on recognition of speech in native and nonnative competing speech. *Journal of Speech, Language, and Hearing Research, 61*(7), 1815–1830. https://doi.org/10.1044/2018_JSLHR-H-17-0254

Gifford, R. H., & Revit, L. J. (2010). Speech perception for adult cochlear implant recipients in a realistic background noise: Effectiveness of preprocessing strategies and external options for improving speech recognition in noise. *Journal of the American Academy of Audiology, 21*(7), 441–451. https://doi.org/10.3766/jaaa.21.7.3

Gifford, R. H., Shallop, J. K., & Peterson, A. M. (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. *Audiology & Neurotology, 13*(3), 193–205. https://doi.org/10.1159/000113510

Herman, R., & Pisoni, D. (2003). Perception of "elliptical speech" following cochlear implantation: Use of broad phonetic categories in speech perception. *The Volta Review, 102*(4), 321–347.

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 and Hearing, 42*(6), 1699–1711. https://doi.org/10.1097/AUD.0000000000001059

Jones, N. P., Siegle, G. J., & Mandell, D. (2015). Motivational and emotional influences on cognitive control in depression: A pupillometry study. *Cognitive, Affective, & Behavioral Neuroscience, 15*(2), 263–275. https://doi.org/10.3758/s13415-014-0323-6

Kamil, R. J., Genther, D. J., & Lin, F. R. (2015). Factors associated with the accuracy of subjective assessments of hearing impairment. *Ear and Hearing, 36*(1), 164–167. https://doi.org/10.1097/AUD.0000000000000075

Kim, S., Wu, Y.-H., Bharadwaj, H. M., & Choi, I. (2021). Effect of noise reduction on cortical speech-in-noise processing and its variance due to individual noise tolerance. *Ear and Hearing, 43*(3), 849–861. https://doi.org/10.1097/AUD.0000000000001144

Konopka, A. E., & Kuchinsky, S. E. (2015). How message similarity shapes the timecourse of sentence formulation. *Journal of Memory and Language, 84,* 1–23. https://doi.org/10.1016/j.jml.2015.04.003

Lai, C. (2010). *What do you mean, you're uncertain?: The interpretation of cue words and rising intonation in dialogue* [Paper presentation]. Proceedings of the Eleventh Annual Conference of the International Speech Communication Association (pp. 1413–1416).

Lemke, U., & Besser, J. (2016). Cognitive load and listening effort: Concepts and age-related considerations. *Ear and Hearing, 37*(Suppl. 1), 77S–84S. https://doi.org/10.1097/AUD.0000000000000304

Lempert, K. M., Chen, Y. L., & Fleming, S. M. (2015). Relating pupil dilation and metacognitive confidence during auditory decision-making. *PLOS ONE, 10*(5), Article e0126588. https://doi.org/10.1371/journal.pone.0126588

Love, J., Sung, W., & Francis, A. L. (2021). Psychophysiological responses to potentially annoying heating, ventilation, and air conditioning noise during mentally demanding work. *The Journal of the Acoustical Society of America, 150*(4), 3149–3163. https://doi.org/10.1121/10.0006383

McCreery, R. W., & Stelmachowicz, P. G. (2013). The effects of limited bandwidth and noise on verbal processing time and word recall in normal-hearing children. *Ear and Hearing, 34*(5), 585–591. https://doi.org/10.1097/AUD.0b013e31828576e2

McGinley, M. J., David, S. V., & McCormick, D. A. (2015). Cortical membrane potential signature of optimal states for sensory signal detection. *Neuron, 87*(1), 179–192. https://doi.org/10.1016/j.neuron.2015.05.038

McHaney, J. R., Tessmer, R., Roark, C. L., & Chandrasekaran, B. (2021). Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry. *Brain and Language, 222,* 105010. https://doi.org/10.1016/j.bandl.2021.105010

McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., & Van Engen, K. J. (2018). Coping with adversity: Individual differences in the perception of noisy and accented speech. *Attention, Perception, & Psychophysics, 80*(6), 1559–1570. https://doi.org/10.3758/s13414-018-1537-4

McLaughlin, D. J., & Van Engen, K. (2020). Task-evoked pupil response for accurately recognized accented speech. *The Journal*

of the Acoustical Society of America, 147(2), EL151–EL156. https://doi.org/10.1121/10.0000718

McLaughlin, S. A., Larson, E., & Lee, A. (2019). Neural switch asymmetry in feature-based auditory attention tasks. *Journal of the Association for Research in Otolaryngology, 20*(2), 205–215. https://doi.org/10.1007/s10162-018-00713-z

McMurray, B., Ellis, T. P., & Apfelbaum, K. S. (2019). How do you deal with uncertainty? Cochlear implant users differ in the dynamics of lexical processing of noncanonical Inputs. *Ear and Hearing, 40*(4), 961–980. https://doi.org/10.1097/AUD.0000000000000681

McMurray, B., Farris-Trimble, A., & Rigler, H. (2017). Waiting for lexical access: Cochlear implants or severely degraded input lead listeners to process speech less incrementally. *Cognition, 169,* 147–164. https://doi.org/10.1016/j.cognition.2017.08.013

Miles, K., McMahon, C., Boisvert, I., Ibrahim, R., de Lissa, P., Graham, P., & Lyxell, B. (2017). Objective assessment of listening effort: Coregistration of pupillometry and EEG. *Trends in Hearing, 21*. https://doi.org/10.1177/2331216517706396

Mirman, D. (2014). *Growth curve analysis and visualization using R*. CRC Press.

Munson, B., Donaldson, G. S., Allen, S. L., Collison, E. A., & Nelson, D. A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *The Journal of the Acoustical Society of America, 113*(2), 925–935. https://doi.org/10.1121/1.1536630

Nieuwland, M. (2021). How 'rational' is semantic prediction? A critique and re-analysis of Delaney-Busch, Morgan, Lau, and Kuperberg (2019). *Cognition, 215,* 104848. https://doi.org/10.1016/j.cognition.2021.104848

O'Neill, E. R., Parke, M. N., Kreft, H. A., & Oxenham, A. J. (2021). Role of semantic context and talker variability in speech perception of cochlear-implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America, 149*(2), 1224–1239. https://doi.org/10.1121/10.0003532

Peelle, J. E., & Van Engen, K. J. (2021). Time stand still: Effects of temporal window selection on eye tracking analysis. *Collabra: Psychology, 7*(1), 25961. https://doi.org/10.1525/collabra.25961

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560–569. https://doi.org/10.1111/j.1469-8986.2009.00947.x

Pisoni, D. B., Broadstock, A., Wucinich, T., Safdar, N., Miller, K., Hernandez, L. R., Vasil, K., Boyce, L., Davies, A., Harris, M. S., Castellanos, I., Xu, H., Kronenberger, W. G., & Moberly, A. C. (2018). Verbal learning and memory after cochlear implantation in postlingually deaf adults: Some new findings with the CVLT-II. *Ear and Hearing, 39*(4), 720–745. https://doi.org/10.1097/AUD.0000000000000530

Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language, 29*(6), 633–654. https://doi.org/10.1016/0749-596X(90)90042-X

Rødvik, A. K., Tvete, O., Torkildsen, J., Wie, O. B., Skaug, I., & Silvola, J. T. (2019). Consonant and vowel confusions in well-performing children and adolescents with cochlear implants, measured by a nonsense syllable repetition test. *Frontiers in Psychology, 10,* 1813. https://doi.org/10.3389/fpsyg.2019.01813

Rönnberg, J., Holmer, E., & Rudner, M. (2019). Cognitive hearing science and ease of language understanding. *International Journal of Audiology, 58*(5), 247–261. https://doi.org/10.1080/14992027.2018.1551631

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*(3), 1017–1031. https://doi.org/10.1016/j.neuroimage.2007.04.066

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a rational and mechanistic account of mental effort. *Annual Review of Neuroscience, 40*(1), 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

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*(4), e14035. https://doi.org/10.1111/psyp.14035

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research, 61*(6), 1463–1486. https://doi.org/10.1044/2018_JSLHR-H-17-0257

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology, 7,* 345. https://doi.org/10.3389/fpsyg.2016.00345

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., & Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hearing Research, 369,* 67–78. https://doi.org/10.1016/j.heares.2018.05.006

Wilson, B. S., & Dorman, M. F. (2008). Cochlear implants: A remarkable past and a brilliant future. *Hearing Research, 242*(1–2), 3–21. https://doi.org/10.1016/j.heares.2008.06.005

Wingfield, A., Alexander, A. H., & Cavigelli, S. (1994). Does memory constrain utilization of top-down information in spoken word recognition? Evidence from normal aging. *Language and Speech, 37*(3), 221–235. https://doi.org/10.1177/002383099403700301

Wingfield, A., Tun, P. A., & Rosen, M. J. (1995). Age differences in veridical and reconstructive recall of syntactically and randomly segmented speech. *The Journal of Gerontology: Series B, 50*(5), P257–P266. https://doi.org/10.1093/geronb/50b.5.p257

Winn, M. B. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends in Hearing, 20,* 1–17. https://doi.org/10.1177/2331216516669723

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The impact of auditory spectral resolution on listening effort revealed by pupil dilation. *Ear and Hearing, 36*(4), e153–e165. https://doi.org/10.1097/AUD.0000000000000145

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 in Hearing, 22,* 1–22. https://doi.org/10.1177/2331216518808962

Winn, M. B., & Teece, K. H. (2020). Slower speaking rate reduces listening effort among listeners with cochlear implants. *Ear and Hearing, 42*(3), 584–595. https://doi.org/10.1097/AUD.0000000000000958

Winn, M. B., & Teece, K. H. (2021). Listening effort is not the same as speech intelligibility score. *Trends in Hearing, 25*. https://doi.org/10.1177/23312165211027688

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 in Hearing, 22*. https://doi.org/10.1177/2331216518800869

Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The pupil dilation response to auditory stimuli: Current state of knowledge. *Trends in Hearing, 22,* 1–25. https://doi.org/10.1177/2331216518777174

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil response as an indication of effortful listening: The influence of sentence intelligibility. *Ear and Hearing, 31*(4), 480–490. https://doi.org/10.1097/AUD.0b013e3181d4f251

Zhang, Y., Lehmann, A., & Deroche, M. (2021). Disentangling listening effort and memory load beyond behavioural evidence: Pupillary response to listening effort during a concurrent memory task. *PLOS ONE, 16*(3), Article e0233251. https://doi.org/10.1371/journal.pone.0233251