Trends in Hearing
Volume 25: 1–26
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/23312165211027688
journals.sagepub.com/home/tia
SAGE

Listening Effort Is Not the Same as Speech Intelligibility Score

Matthew B. Winn and Katherine H. Teece

Abstract

Listening effort is a valuable and important notion to measure because it is among the primary complaints of people with hearing loss. It is tempting and intuitive to accept speech intelligibility scores as a proxy for listening effort, but this link is likely oversimplified and lacks actionable explanatory power. This study was conducted to explain the mechanisms of listening effort that are not captured by intelligibility scores, using sentence-repetition tasks where specific kinds of mistakes were prospectively planned or analyzed retrospectively. Effort measured as changes in pupil size among 20 listeners with normal hearing and 19 listeners with cochlear implants. Experiment 1 demonstrates that mental correction of misperceived words increases effort even when responses are correct. Experiment 2 shows that for incorrect responses, listening effort is not a function of the proportion of words correct but is rather driven by the types of errors, position of errors within a sentence, and the need to resolve ambiguity, reflecting how easily the listener can make sense of a perception. A simple taxonomy of error types is provided that is both intuitive and consistent with data from these two experiments. The diversity of errors in these experiments implies that speech perception tasks can be designed prospectively to elicit the mistakes that are more closely linked with effort. Although mental corrective action and number of mistakes can scale together in many experiments, it is possible to dissociate them to advance toward a more explanatory (rather than correlational) account of listening effort.

Keywords

listening effort, speech intelligibility, speech perception, cochlear implants, pupillometry

Received 17 August 2020; Revised 16 April 2021; accepted 28 May 2021

To better understanding hearing loss, listening effort should be examined with equal if not greater emphasis than speech intelligibility. Hughes et al. (2018) described in detail the reports of hearing-impaired individuals who frequently commented on burdensome effort leading to withdrawal from social communication because of anxiety and embarrassment but who did not apparently complain of low speech intelligibility per se. A likely explanation is that sentences are not misunderstood when the listener has simply chosen to avoid the conversation altogether. Apart from the difficulties described by Hughes et al., listening effort is also thought to lead to mental fatigue and increased stress (Hétu et al., 1988), unemployment (Järvelin et al., 1997), early retirement (Danermark & Gellerstedt, 2004), need for recovery time after work (Nachtegaal et al., 2009), and emotional strain (Alhanbali et al., 2018).

It is tempting and intuitive to accept speech intelligibility scores as a proxy for listening effort, but this link is likely oversimplified and lacks actionable explanatory power. As an analogy, hearing thresholds are correlated with age, but age is not a sufficient proxy for setting acoustic gain for a hearing aid. Similarly, correlations between effort and intelligibility—if they do arise—do not explain what makes speech effortful. The current study tracks changes in effort in speech perception that are linked with the types, numbers, and positions of intelligibility

Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, United States

Corresponding author:
Matthew B. Winn, University of Minnesota, Twin Cities, 164 Pillsbury Dr SE, Minneapolis, MN Minnesota 55455, United States.
Email: mwinn@umn.edu

errors, to explore whether these factors are a better explanation of effort than the proportion of words correct.

The common assumption that individuals with poorer speech intelligibility are the ones suffering with greater effort has not been established empirically. Zekveld et al. (2010) demonstrated a relationship between intelligibility and pupil-related signs of effort in listeners with normal hearing (NH), but Ohlenforst et al. (2017) and Wendt et al. (2018) clarified that this relationship is nonmonotonic and invited speculation that individuals could avoid investing effort in very difficult situations (cf. Eckert et al., 2016; Winn et al., 2018). Setting aside the notion that effort decreases when a listener gives up, there could still be questions about the relationship between effort and intelligibility even in the performance range where people are still trying. In other words, it is already known that intelligibility and effort have a nonmonotonic relationship, but the link deserves further exploration. A person's reported effort can be entangled with their own sense of their performance in unpredictable ways. Moore and Picou (2018) have shown that individuals report self-assessed performance (i.e., accuracy) when asked to rate their own listening effort and that these two constructs can be anticorrelated. In line with that, O'Neill (2020) has shown that people with cochlear implants (CIs) can sometimes rate their performance as *not difficult* even when their intelligibility scores were in the range of 40%–60%. These cases provide further evidence that effort does not scale linearly with intelligibility and should be measured separately.

An additional complication of the effort of speech perception is that speech unfolds rapidly over time, with cognitive processing tied to specific landmarks in a stimulus on the order of hundreds of milliseconds (Altmann & Kamide, 1999; Altmann & Mirković, 2009). The effort associated with speech perception therefore might not be feasible to track subjectively in real time *during* the relevant perceptual events. Similarly, intelligibility scores are an *after-the-fact* one-time measurement that fails to indicate the processing that led up to a final answer. Just as eye-tracking studies of spoken word recognition have revealed the nature of incremental perception and perceptual competition among words (Altmann & Kamide, 1999), it is the series of cognitive activity that occurs between the perception and the response that will potentially shed light on the mental effort experienced when understanding speech.

The Invisible Problem of Mental Correction

In speech recognition tasks, listeners can safely assume that they will hear coherent and sensible utterances and will tend to repeat back coherent responses, with their guesses guided by the knowledge and constraints of the language—what Rönnberg et al. (2019) call *postdiction*. Even when stimulus materials are designed to specifically lack linguistic structure, listeners will still process meaning (Ivanova et al., 2012) and will still impose structure in their verbal responses both semantically (Popa, 2018) and phonetically (Herman & Pisoni, 2003). For individuals with hearing loss, this invisible mental transformation is obviously an adaptive skill that helps them carry on in conversation. But for audiologists—or anyone interested in measuring speech perception (hearing scientists, engineers, psychologists, etc.), this is an insidious problem, because this correction process is likely the source of the mental energy we seek to quantify, yet there is no obvious way to detect whether it has happened, since it can result in an unremarkable correct response.

The current study focuses on the problem of mental correction in speech perception because of its relevance to understanding the experience of people with hearing loss, who are more likely to misperceive words and therefore more likely to have their effort driven by mental correction of those errors. We expect that speech perception usually involves simplified heuristics to quickly predict and resolve perceptions without excessive deliberation. When the expected relationships between words are violated, active intentional cognitive control can be exerted, bounded by the limited resources available (Griffiths et al., 2015) and also bounded by the quick rate of information transmission in speech. It is important to recognize that not all speech perception mistakes would result in a violation of linguistic coherence. Therefore, there are opportunities to dissociate effort and an intelligibility score by separately tracking the number of errors and the linguistic incoherence.

Classifying Types of Responses in Speech Intelligibility Tests

Traditional intelligibility scoring treats all errors as linearly additive; each error is worth the same as each other error. This principle simplifies tallying scores and describing outcome measures, but it overlooks the potential unequal weighting of errors as they contribute to effort. Ultimately, a better understanding of the factors that drive listening effort can guide the development of evaluation materials (i.e., perception tests) that target those specific relevant factors rather than testing for intelligibility in a way that is agnostic to—or worse—*uncorrelated with*—listening effort.

As a prelude to the analysis to come, we offer a simple taxonomy of response/error patterns, with examples displayed in Table 1. Response Type 1 is ideal, when the

Table 1. Classification of Specific Response Types in Sentence-Perception Experiments.

Type	Description	Example stimulus	Example response
1	Correct perception, correct response		
2	Misperceived word correctly filled in by context	She xxxx the candle with a match	She *lit* the candle with a match
		The bread is made from whole wheat	The *bird* is . . . The *bread* is made from whole wheat
3	Semantically sensible replacement of missing word	The doctor prescribed the drug	The doctor prescribed the *pill*
		She used a cloth to dry the dishes	She used a *towel* to dry the dishes
4	Similar-sounding replacement of missing word; sentence is sensible	The most important factor is the clock	The most important factor is the *plot*
		Airplanes require a special staff	*Airlines* require a special staff
5	Similar-sounding replacement of word; sentence is nonsense	I can't guess, so give me a clue	I *cut this*, so give me a clue
		The glass had a chip on the rim	The glass had a *check* with the *wind*
6	Multiple mistakes to accommodate one mistake	She made the bed with clean sheets	She made the *bagel* with *cream cheese*
		They were considering the cheers	They were *sitting* in the *chairs*

listener hears all the words correctly and repeats the words correctly. Type 2 is when one or more of the words were not perceived correctly, but the context of the sentence is sufficient for the listener to correctly guess. Experienced users of a language can correctly fill in a missing word based on context or intuition, disguising the fact that the word was not heard correctly, or perhaps not heard at all. Response Types 1 and 2 are indistinguishable on intelligibility tests, even though they are likely unequal in terms of effort, since Type 2 involves extra restorative processing. Explicitly acknowledging response Types 1 and 2 as distinct will ultimately help move toward a mechanistic account of effort that also relates to correct responses rather than just incorrect responses. That distinction is explored in the current study in Experiment 1.

Response Type 3 is when a listener repeats a word in a sentence incorrectly, but the word is a sensible replacement within the context and is also acoustically similar. For example, the stimulus *Airlines require a special staff* repeated as *Airlines require a special stash* or *Airlines require a special trash*. Response Type 4 is when the listener incorrectly substitutes a word that is sensible but is acoustically dissimilar to the actual word (e.g., the participant's response is *Airlines require a special pilot*), which is a clear example of the influence of language overriding the input of the auditory system. These errors invite the listener's confidence in a false perception because they are coherent, so they might not necessarily evoke a substantial increase in effort. If effort is related to the mental activity of substituting one word with another related word, then error Types 2, 3, and 4 could be similar in terms of effort but would produce different intelligibility scores, at both the word and phoneme levels.

Response Type 5 is when a word is heard incorrectly, and the listener's response is not linguistically coherent or sensible. For example, the sentence *The glass had a chip on the rim* repeated as *The glass had a check with the wind* is a Type 5 response. We suspect that Type 5 elicits the greatest effort, since the listener is likely to engage in some mental activity to ponder how the sentence can be altered to make sense.

Error Type 6 is arguably the most interesting—it is when a listener hears a word incorrectly, and then a *different* part of the sentence is incorrectly transformed to accommodate that error, perhaps because the listener is unaware that an error was made. For example, *Airlines require a special staff* could be repeated as *Air mail requires a special stamp*; both incorrect syllables are coherent with each other, rendering the response sensible, even if incorrect. Coherent responses can result from a larger number of errors as well. For example, *She made the bed with clean sheets* was repeated as *She made the bagel with cream cheese*. This error could result from hearing *sheets* as *cheese* or from hearing *bed* as *bagel*. Either of these single errors could cause a post hoc reworking of the sentence to agree with the mistaken word. From the experimenter's point of view, this response has multiple errors, and none of the meaning of the actual target sentence was retained; it is a failed perception. But from the listener's perspective, the response is more sensible than *She made the bed with clean cheese*, even though that nonsense response would score *higher* on a test of intelligibility since it contains one error instead of three. Setting aside the possibility that the listener genuinely misperceived *three* words that were all semantically coherent with each other by chance, it is likely that these error patterns arise because of linguistic constraints or postdiction. These situations are crucial for the current investigation because they provide the test of whether the effort of speech understanding scales with linguistic coherence or the total number of errors. If intelligibility errors were the source of effort, then the three-error sentence should be more effortful. But if the linguistic coherence is the source of effort, then the single-error incoherent sentence should be more effortful. This hypothesis is tested in the current study in Experiment 2.

Measuring Listening Effort Using Pupil Dilation

Task-evoked pupil responses have been used in a wide variety of cognitive tasks (Beatty, 1982; Laeng et al., 2012) including numerous studies of speech perception (see Zekveld et al., 2018 for review). The general trend is to observe greater pupil dilation in cases where more effort is exerted, although the measurement can be complicated and nuanced (Winn et al., 2018). Physiological studies of pupil dilation have suggested its link to the locus coeruleus (Aston-Jones & Cohen, 2005; Murphy et al., 2011) and broadly distributed cortical activity (Reimer et al., 2016). Pupil size has been shown to predict dynamic cortical states that govern task performance (McGinley et al., 2015), underscoring the role of the cingulo-opercular neural network, which is involved in monitoring errors and resolving uncertainty in perception (Eckert et al., 2016; Vaden et al., 2013, 2017). Although the physiological link between pupil dilation and the cingulo-opercular network is yet to be fully described in the literature, studies that coregister both measurements suggest coupled activity (Breeden et al., 2017; Kuchinsky et al., 2016). Kuipers and Thierry (2011) simultaneously measured pupil dilation and electroencephalogram (EEG) responses evoked by violating semantic predictions, further highlighting the potential for focused neural localization of the process under investigation here.

The current study is not the first to examine changes in pupil dilation linked to aspects of speech perception other than intelligibility. There are systematic differences in pupil size elicited by stimuli that vary by factors that are more subtle than all-or-none correctness, such as lexical competition (Kuchinsky et al., 2013), speech accentedness (McLaughlin & Van Engen, 2020), speaking style (e.g., conversational versus clear; Borghini & Hazan, 2020), sentence structure (Ayasse & Wingfield, 2018; Demberg & Sayeed, 2016), translation from a different language (Hyönä et al., 1995), pronoun resolution (Vogelzang et al., 2016), semantic context (Borghini & Hazan, 2020; Winn, 2016), and lexical ambiguity (Kadem et al., 2020). We therefore expect that different types of perceptual patterns should elicit measurable differences in listening effort reflected in pupil dilation, in a granular fashion, independent of intelligibility scores.

Summary and Hypotheses

Gathering the ideas described earlier, a simple tally of errors will inevitably lose the rich information about the structure of the misperception and the work that would be needed to make sense of it. The results of previous literature do not disentangle the two competing hypotheses of listening effort being driven by the number of mistakes versus the degree of linguistic ambiguity/coherence, since the situations that demand mental disambiguation are likely to also be the situations that result in a greater number of mistakes. Using a combination of prospective experimental design as well as post hoc analyses of previous data, the current study uses two data sets with a sufficiently high number of errors and a sufficient diversity of errors to map the effort associated with various patterns of intelligibility. The taxonomy of response/error types described earlier leads to the following hypotheses:

1. Even if a response is correct, more effort will be exerted if a listener needs to resolve an ambiguity to generate a meaningful response (response Type 2 elicits greater effort than response Type 1).
2. When a response is incorrect, listening effort will be only slightly elevated if the incorrect word is sensible in its context (Types 3 and 4 will elicit similar to effort for response Type 2 and only slightly higher than Type 1).
3. A single error could be more effortful than multiple errors, if that single error leads to a linguistically incoherent sentence (Type 5) and if the multiple errors lead to a coherent/sensible sentence (Type 6). That is, we expect no main effect of the number of errors, but we expect an effect of linguistic coherence.
4. Based on the important role of predictive processing in speech perception, errors that occur earlier in a sentence should be more costly than errors that occur later in a sentence.

Taken together, these statements collectively hypothesize a double dissociation between listening effort and intelligibility, and they propose a more specific association between listening effort and resolution of linguistic ambiguity. Hypothesis A is tested by Experiment 1, and the other hypotheses are tested by Experiment 2.

Experiment 1: Elevated Effort for Correct Responses

Correct responses do not clearly convey whether a sentence was heard correctly, or if it was heard incorrectly and then mentally corrected before the listener gave their response. We expect that the need for mental correction is effortful, but the phenomenon itself is difficult to verify empirically (i.e., it is the *invisible* problem described in the introduction). To address this challenge, the first experiment used a new stimulus set designed specifically to demand retroactive perceptual restoration of a missing word, but not in a way that would affect intelligibility in any meaningful way. If effort were elevated in such cases, it would partly disentangle effort from intelligibility by showing that perception could be effortful despite perfect scores. This approach tests the hypothesis by Rönnberg et al. (2019) that postdiction—reconstruction of fragmented perceptions—elicits increased effort.

Experiment 1: Methods

Participants. There were 20 young adult listeners with NH thresholds, who all spoke English as their native language and who reported no language or learning disabilities. Participants were not evaluated for visual acuity. All gave informed written consent of procedures that were approved by the Institutional Review Board at the University of Minnesota.

Stimuli. Stimuli included 120 sentences, each of which was designed to have a target word early in the sentence (2nd, 3rd, or 4th word) that was not predictable based on preceding words but was narrowly constrained based on subsequent words, for example, *The woman ___ her candle with a match*, where the target word is *lit*. The sentences had an average of nine words, and the target word occurred on average at word position 3.35. 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. Each sentence was spoken multiple times by a trained audiologist (the first author) with the explicit intention of being clear and slow enough to promote excellent intelligibility even among people with hearing loss (for a future study). The recordings were examined by a team of five other listeners who are audiologists or AuD students, and then passed on to the subsequent stages of processing, described later.

To verify the contextual constraint on the target words, an online test was conducted using the Gorilla experiment builder (www.gorilla.sc; Anwyl-Irvine et al., 2019) where 27 participants were presented with the written sentences with the target word missing and were asked to type in the missing word. The proportion of participants who give the same response (ignoring changes in tense and plurality) is the *cloze probability* (Kutas & Hillyard, 1984). There are varying criteria for what is considered *high* and *low* cloze probability, but Block and Baldwin (2010) suggest a *high* probability answer to be at 67% and above. Using this metric, there was high cloze probability for 100 of our 120 sentences, with low (< 33%) probability for just one sentence.

Stimulus Variations. There were three versions of each sentence, illustrated in Figure 1. The *intact* version was the plain utterance with all words spoken naturally. There were two versions that distorted the target word, which forced the listener to engage in some perceptual

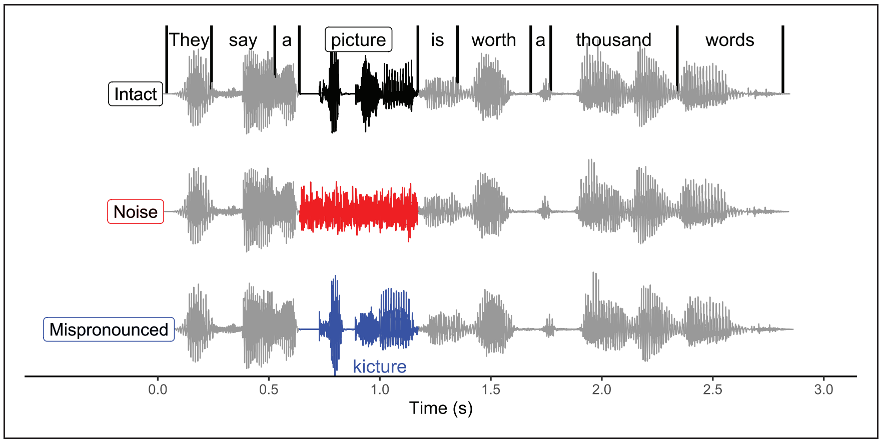

Figure 1. Types of stimuli used in the sentence recognition experiment, including an “intact” form (top) where all the words were unaltered, a “noise” form (middle) where a single target word was replaced with noise of equal duration and intensity, and a “mispronounced” form (bottom) where a phoneme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

restoration. In the *Noise* condition, the target word was replaced with noise of equivalent 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 (Bilger & Wang, 1976; Dubno et al., 1982), including those who wear CIs (Munson et al., 2003; Rødvik et al., 2019). The mispronunciations were intended to be nonwords, but 18 (out of 120) of the mispronunciations produced real English words. The sentences with mispronunciations were spoken with the same prosody by the same talker. The recording of the sentence up to and including the mispronounced word was spliced onto the intact form of the sentence starting from the end of the target word. The effect of this splicing was to ensure that the audio content following the target word—which served to disambiguate the target word itself—was exactly the same in all versions of the stimuli.

Procedure. 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 lists of 30 sentences. Each list began with an intact sentence, followed by a random ordering of stimulus types, with the stipulation that the same stimulus type could not appear in more than three consecutive trials. 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. Therefore, 80 of the 120 of the sentences contained a word to be disambiguated; however, the ambiguity of the word itself was extremely low, as verified by the cloze test.

During the experiment, listeners sat in a chair with their head position stabilized by a chin rest. 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 prompt 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.

During the task, the participant's eye position and pupil size were monitored and 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 to minimize influence of the pupillary light reflex.

Experiment 1: Analysis

Intelligibility. Intelligibility for each word in the sentence was scored in real time by an experimenter, and the participant's verbal responses were audio recorded for follow-up inspection and verification. For mispronounced stimuli, any word that did not rhyme with the mispronounced target word was counted as an error, as well as any errors elsewhere in the sentence. For stimuli where the target was replaced by noise, errors included any word that was not semantically coherent with the stimulus, as well as any errors elsewhere in the sentence. If the participant's guess at the word replaced by noise was not the "intact" 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.

Pupillometry Data Preprocessing. Pupil data were processed in the style described by Winn et al. (2018). 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 processed with a 5 Hz low-pass 4th-order Butterworth filter and then decimated to sample once every 40 ms. Baseline pupil size was calculated in time spanning 500 ms before stimulus onset to 500 ms after stimulus onset, and each pupil size data point in the trial was expressed as a proportional difference from the trial-level baseline. Other methods of baselining (e.g., linear subtraction) have been proposed, with debate over which is the best option (Mathôt et al., 2018; Reilly et al., 2019; Winn et al., 2018). Reilly et al. (2019) suggests that the pupillary response is linearly consistent across varying baseline pupil size, though their study altered baseline size using luminance changes whose effects have not been verified to generalize to other baseline influences such as arousal. Pupil data were tagged with timestamps to enable aggregation by onset or offset of the sentence, as well as onset/offset of the target word.

Pupil data were subjected to an algorithm written to automatically detect and remove potential contaminations or outliers in the morphology. There were several rules to "flag" potential contamination, such as to detect hippus activity that could contaminate baseline correction for the whole trial (e.g., baselines that unusually deviated from both the previous and the next baseline, a baseline that deviated from the mean baseline by at least 2 standard deviations, a significant slope of change in pupil size during the baseline), a steep downward slope of pupil size immediately after stimulus onset (which, when combined with the baseline flags, would confirm contamination stimulus-unrelated dilation before the trial). Apart from baseline flags, there were also flags for baseline-proportioned pupil size ±3 standard deviations from the mean, or absolute pupil size 3 standard deviations from the mean. Trials were excluded if they contained three or more flags, or if they contained 30% missing data during the time period from the start of the baseline leading to 3 s past the onset of the stimulus. Ten percent of trials were discarded due to these criteria. Of the discarded trials, 42% were from the Intact condition, 34% from the Mispronounced condition, and 24% from the Noise condition.

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 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 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 stimulus offset, which corresponded to the *listening* phase of each trial. Window 2 spanned from 0.7 to 2.2 s relative to stimulus offset, reflecting the *response preparation* phase of the trial. These windows arguably correspond to auditory encoding versus poststimulus linguistic resolution and 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 et al., 2015; Winn & Moore, 2018; Winn & Teece, 2020).

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:

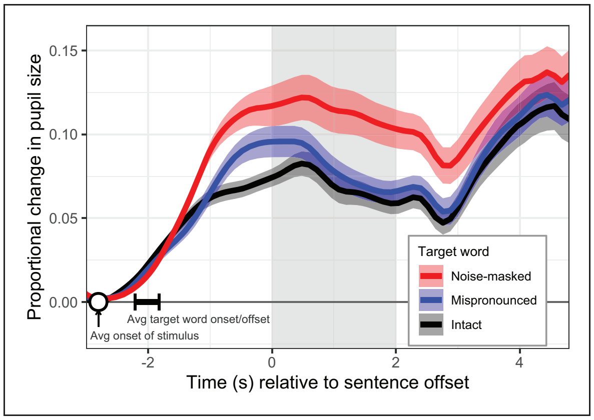

Figure 2. Proportional Change of Pupil Dilation Over Time in Response to Stimuli Where an Early-Occurring Target Word Was Intact (Black), Replaced by Noise (Red), or Intentionally Mispronounced (Blue). Width of the ribbon around the data represents ±1 standard error. The gray shaded region corresponds to the 2 s that elapsed between stimulus offset and the response prompt. The average position of the target word is indicated by the horizontal *error bar* along the *x*-axis, which also indicated the average stimulus onset time in an open circle.

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. Deidentified data are available upon request.

Experiment 1: Results

As expected, intelligibility errors were extremely rare, with performance at 99.2%. Figure 2 illustrates the changes in pupil dilation in response to the three types of stimuli. The summary of Experiment 1 statistical analysis of the pupil data is in Table 2 (Window 1) and Table 3 (Window 2), with the Mispronounced condition set as the model default, to allow comparisons in both the upward and downward direction. During the first analysis window, stimuli with intact target words elicited less overall pupil dilation than those with mispronounced words (Table 2, $\beta 4$, $t = –2.72$, $p = .015$), and Noise-masked targets elicited more pupil dilation compared to mispronounced words ($\beta 5$).

There was shallower growth of pupil size for Intact sentences compared to the Mispronounced stimuli, reflected in the linear slope term ($\beta 6$). The detectable

```
lmerTest :: lmer(pupil ~ poly1 + poly2 + Type +      # Main effects
poly1 : Type  +  poly2 : Type +                      # two – way interactions
(1 + poly1 + poly2 + Type+                           # main random effects
poly1 : Type  +  poly2 : Type | Listener),           # random interactions
data  =  Data_window)
```

Table 2. Linear Mixed-Effects Model Accounting for Change in Pupil Size in Experiment 1 for Window 1 (–1.5 to 0.7 s Relative to Stimulus Offset).

	Term	Estimate	*SE*	*df*	*t*	*p*
*β*1	Intercept (Mispronounced)	0.085	0.007	16.00	11.72	<.001
*β*2	Linear	0.046	0.013	16.01	3.58	.003
*β*3	Quadratic	–0.028	0.004	16.00	–7.28	<.001
*β*4	Type-Intact	–0.015	0.005	16.00	–2.72	.015
*β*5	Type-Noise	0.024	0.005	16.00	4.79	<.001
*β*6	Linear:Type-Intact	–0.016	0.007	16.00	–2.14	.048
*β*7	Linear:Type-Noise	0.004	0.011	15.99	0.34	.737
*β*8	Quadratic:Type-Intact	0.029	0.006	16.00	4.71	<.001
*β*9	Quadratic:Type-Noise	0.005	0.005	15.99	1.09	.290

Window 1 Random Effect Variance-Covariance Matrix				Correlations							
		Var	*SD*	*β*1	*β*2	*β*3	*β*4	*β*5	*β*6	*β*7	*β*8
*β*1	Intercept (MP)	0.0009	0.030								
*β*2	Linear	0.0028	0.053	0.47							
*β*3	Quadratic	0.0002	0.016	–0.84	–0.34						
*β*4	Type-Intact	0.0005	0.022	–0.74	–0.30	0.67					
*β*5	Type-Noise	0.0004	0.020	0.13	0.33	–0.21	0.20				
*β*6	Linear: Type-Intact	0.0009	0.030	–0.01	–0.65	–0.21	–0.12	0.01			
*β*7	Linear: Type-Noise	0.0021	0.046	0.60	–0.12	–0.55	–0.41	0.30	0.49		
*β*8	Quadratic: Type-Intact	0.0006	0.024	0.64	0.47	–0.79	–0.59	0.19	–0.10	0.21	
*β*9	Quadratic: Type-Noise	0.0004	0.020	0.20	–0.18	–0.49	–0.50	–0.35	0.43	0.06	0.33

Note. SE = standard error of the mean estimation; *df* = degrees of freedom estimated using the Satterthwaite approximation (implementation by Kuznetsova et al. 2017); Var = Variance; *SD* = standard deviation; MP = mispronunciation condition (model default).

Table 3. Linear Mixed-Effects Model Accounting for Change in Pupil Size in Experiment 1 for Window 2 (0.7 to 2.2 s Relative to Stimulus Offset).

	Term	Estimate	*SE*	*df*	*t*	*p*
*β*10	Intercept (Mispronounced)	0.072	0.007	16.09	10.03	<.001
*β*11	Linear	–0.023	0.010	15.98	–2.35	.032
*β*12	Quadratic	0.011	0.002	15.99	4.42	<.001
*β*13	Type-Intact	–0.007	0.003	16.02	–2.41	.028
*β*14	Type-Noise	0.038	0.007	16.06	5.48	<.001
*β*15	Linear:Type-Intact	0.003	0.005	15.94	0.61	.554
*β*16	Linear:Type-Noise	0.003	0.005	16.00	0.59	.565
*β*17	Quadratic:Type-Intact	–0.004	0.002	15.96	–1.55	.140
*β*18	Quadratic:Type-Noise	–0.011	0.002	15.93	–6.00	<.001

Window 2 Random Effect Variance-Covariance Matrix				Correlations							
		Var	*SD*	*β*10	*β*11	*β*12	*β*13	*β*14	*β*15	*β*16	*β*17
*β*10	Intercept (MP)	0.0009	0.030								
*β*11	Linear	0.0016	0.040	–0.33							
*β*12	Quadratic	0.0001	0.009	–0.33	–0.14						
*β*13	Type-Intact	0.0001	0.012	–0.32	0	–0.20					
*β*14	Type-Noise	0.0008	0.029	0.52	–0.58	–0.12	–0.17				
*β*15	Linear: Type-Intact	0.0004	0.021	0.17	–0.61	0.32	0.01	0.21			
*β*16	Linear: Type-Noise	0.0005	0.022	0.37	–0.38	–0.12	0.02	0.15	0.39		
*β*17	Quadratic: Type-Intact	0.0001	0.009	0	0.26	–0.23	–0.05	0.15	–0.52	0.06	
*β*18	Quadratic: Type-Noise	0.0000	0.006	0.07	0.59	–0.22	–0.36	–0.19	–0.24	0.10	0.6

Note. SE = standard error of the mean estimation; *df* = degrees of freedom estimated using the Satterthwaite approximation (implementation by Kuznetsova et al. 2017); Var = Variance; *SD* = standard deviation; MP = mispronunciation condition (model default).

curvature down from the peak was detectable for Mispronounced targets, reflected in the quadratic term ($\beta3$), and this curvature was not different for sentences with noise-masked targets ($\beta9$). However, this curvature was completely neutralized for the Intact stimuli, as indicated by an interaction of the same degree but opposite direction ($\beta8$). This result is likely due to the lack of elevated peak for the Intact stimuli, thus reducing the potential curvature down from that peak.

During the second analysis window, overall pupil dilation (the intercept term) was statistically smaller for Intact stimuli compared to dilation for Mispronounced stimuli (Table 3, $\beta13$, $t = –2.41$, $p = .028$). Noise-masked stimuli elicited the largest pupil dilation, which was statistically larger than dilation for the Mispronounced stimuli ($\beta14$, $t = 5.48$, $p < .001$). There were no statistical differences between conditions for the linear slope term in the second window. Curvature in the dilation response was detectable for the Mispronounced stimuli, as reflected by the quadratic term ($\beta12$) and that curvature was not different for the Intact stimuli ($\beta17$). However, that curvature for the Noise-masked targets was reduced by a negative interaction of the quadratic term ($\beta18$). As a result of that interaction, the summed quadratic term for Noise-masked targets was effectively zero. This pattern of results for the quadratic term refers to the larger downward inflection in pupil size between the sentence and the verbal response in the Intact and Mispronounced-Target stimulus conditions, compared to the Noise condition in which the large dilation persisted across the time interval between stimulus and response.

Experiment 1: Discussion

The main pattern of results in Experiment 1 support Hypothesis A: The need for mental correction will elicit greater effort even when a response is correct. Within the framework of responses described in the Introduction, the *Intact* responses correspond to Type 1, and the *Mispronounced* and *Noise-Masked* responses both correspond to Type 2. Within this response type, changes in pupil size appear to scale with the amount of perceptual restoration needed, with more dilation for restoring a whole word compared to restoring one phoneme. The earlier onset of dilation for the Noise stimuli likely reflects the difference in mere *identification* of a distortion (which is immediately evident for noise-masked words) versus *processing* a linguistically inappropriate pronunciation. However, at least part of the elevated response for sentences with masked words could be attributable to the mere detection of an acoustic anomaly in the sentence. But this is unlikely to be the sole effect, given the long time course of elevated dilation that persisted through the entire sentence, through the retention interval and into the verbal response (compared to explicit detection tasks which elicit dilations that last less than 1 s; Beatty, 1982; McCloy et al., 2016).

In normal testing circumstances, it is not possible to know for sure whether a listener engaged in perceptual restoration when giving a correct response, but in the current study, the need for retroactive restoration was built into stimulus design. This ensured that the correct responses resulted from post hoc restoration rather than genuine correct perception or prediction. Previous studies of perceptual restoration (e.g., Bhargava et al., 2014; Warren, 1970) have also used speech interrupted by periodic noise or selective masking of words, but without the specificity of targeting specific words or ruling out predictive use of context. Typically, mistakes resulting from masking noise, vocoding, reverberation, and so forth are mistakes outside the experimenter's control; the mistakes could be different for different listeners and different for random iterations of the same stimulus. Conversely, each perceptual restoration in the current study was prospectively designed so that it could be timestamped and linked with a particular change in effort and a specific context for disambiguation.

Two-thirds of the stimuli in Experiment 1 required the listener to correct something about the perception before submitting their response. It is therefore possible that the expectation of the need for active cognitive control elicited higher-than-normal engagement for the NH listeners who were tested. Expectations of prolonged or elevated auditory processing could influence the growth of pupil dilation (McCloy et al., 2017; Winn & Moore, 2018). However, the stimulus types in the current experiment were randomized rather than blocked and were not signaled before the trial. Therefore, the need for mental correction was not predictable at the moment of perception. Despite this randomization, there was a carryover effect of stimulus type between trials, as Intact sentences elicited slightly higher pupil dilation following sentences with noise-masked target words compared to the other conditions, and sentences with noise-masked target words elicited smaller pupil dilation when following sentences that also had noise-masked target words (see Supplementary Material 1, and cf. Vaden et al., 2013 for a discussion of this phenomenon). Further exploring details of the data, we found that using a subtractive baseline correction resulted in a pattern of data that was virtually indistinguishable from the data shown here, which used a divisive baseline procedure (see Supplemental Material 2).

Experiment 2: Effort Patterns Associated With Various Types of Correct and Incorrect Responses

A previous study (Winn & Teece, 2020) involving CI participants listening to Revised Speech-in-Noise

(R-SPiN; Bilger et al., 1984) sentences was reexamined for the relative frequency of occurrence of various error types and was found to be a rich source of data to explore the main issue outlined in the Introduction. Specifically, there were a substantial number of linguistically coherent and incoherent error types that could complement the types of responses in Experiment 1 (which had only Type 1 and Type 2 responses) to fill out the taxonomy described in the Introduction. The R-SPiN corpus is ideal for this investigation because it contains both high-predictability and low-predictability target words in sentence-final position, thus allowing context to either supplement or override the acoustics and to promote cognitive processes that demonstrate when semantic coherence played a role in the participant's response.

Experiment 2: Methods

The participants, stimuli, and procedure were described previously by Winn and Teece (2020), and this is a reexamination of their data. The basic sentence-repetition testing protocol was the same as for Experiment 1 described earlier, with different test stimuli and different participants.

Participants. Data were collected from 21 adults with CIs (age range: 23–82 years, average: 61). Two were excluded from data analysis because of poor camera tracking or excessive data loss. All participants were native speakers of North American English. All participants were able to converse freely during face-to-face communication, and none reported cognitive or language-learning difficulties. All but one participant acquired hearing loss after language acquisition; the sole peri-lingually deafened individual had very good speech intelligibility and was deemed capable of performing well enough to be included in the group. However, as would be expected by their use of CIs, they exhibited a wide range of perceptual errors during auditory-only speech perception that enabled the current analysis, but which would be less likely to emerge in a study of listeners with NH. The median length of CI use was 6 years, with a range of 1–28 years. Of the 19 participants whose data were used, 12 were bilaterally implanted and 7 were unilaterally implanted. Two participants routinely wore a hearing aid in the ear contralateral to unilateral implantation. All were tested using their everyday listening settings, except that the participants with hearing aids were asked to remove the aids during testing; one of these participants preferred to use her hearing aid during testing and was permitted to do so. Participants were not evaluated for visual acuity; those who wore glasses typically removed them for the test, as there was no visual stimulus of consequence apart from the basic color change from red to green.

Stimuli. Stimuli were a subset of the R-SPIN materials (Bilger et al., 1984) used previously by Winn and Moore (2018) and Winn and Teece (2020). The subset was selected to avoid examples of outdated language and to avoid emotional or evocative language that might disproportionately influence pupil dilation measurements. There was a total of 114 high-context and 118 low-context sentences. The final word in each sentence is considered to be the *target* word. A high-context sentence such as *The lion gave an angry roar* gives the listener the chance to predict the target, but a low-context sentence such as *They thought about the roar* provides no help in anticipating the target word.

Procedure. The procedure was the same as for Experiment 1. Each testing session began with a set of five sentences to familiarize the listener with the pace of the test and the style of sentences that they would hear. Following the practice, there were four blocks of 29 sentences each.

Analysis. Spoken responses were analyzed by the authors to tag errors with several nonexclusive labels, including errors based on phonetics (with subgroups of onsets, vowels, and offsets), semantic coherence, syntactic structure, names, articles and pronouns, guesses that were unrelated to the stimulus, guesses that were linguistically incoherent, and guesses at words that were arguably more plausible than the ones in the actual target sentence. For each trial, one author scored each sentence, and then the other author cross-checked the error. After each error was marked, the authors then convened with the full data set to finalize the organization. Table 4 contains examples of the various error types.

Among all of the listeners, there were a total of 1,898 trials, out of which 1,766 trials (93%) had viable pupil data. There were 1,207 trials with correct responses, and 559 trials with at least one error. Some sentences contained multiple errors, and some errors were given multiple tags. For example, the sentence *They were considering the cheers* repeated as *They were sitting in the chairs* was tagged as having a phonetic error (cheers/chairs), a semantic coherence error (sitting, related to chairs) and also being more plausible than the original sentence. In addition, we kept a simpler tally of the mere presence or absence of an error on the sentence-final target word, and the presence or absence of an error on any of the words leading up to the target word (as explained in previous papers by Winn, 2016; Winn & Moore, 2018; Winn & Teece, 2020). There was also a tally of whether a target word error was *accompanied* by an error on the leadup words, and vice versa.

Table 4. Names and Examples of Error Tags.

Phonetic	
Stimulus	*Participant response*
You knew about the clip	You knew about the **click**
The ship's captain summoned his crew	The **shift** captain summoned his crew
Semantic	
The bride wore a white veil	The bride wore a white **gown**
We heard the ticking of the clock	We heard the **chicken** and the **cluck**
Phonetic and Semantic	
We are considering the cheers	We are **sitting** in the **chairs**
We shipped the furniture by truck	We **sent** the furniture by truck
Syntax	
Tighten the belt by a notch	**I'm thinking about buying** a watch
The swimmer dove into the pool	**Swim with Doug** into the pool
Segmentation	
A bicycle has two wheels	The **bus only** has two wheels
The dealer shuffled the cards	The **dealership** sold the cards
Article/pronoun	
We heard the ticking of the clock	**He** heard the ticking of the clock
Her entry should win first prize	**The** entry should win first prize
Total guess	
The fruit was shipped in wooden crates	The **machines** were shipped in wooden crates
Ruth poured the water down the drain	Ruth **paddled across the bay**
More plausible	
He's glad you called about the jar	He's glad you **could open** the jar
We are considering the sheers	We are considering the **choice**
Nonsense	
Mary wore her hair in braids	**May what we get in days**
The glass had a chip on the rim	The glass had a **check with the wind**

Note: Bold text refers to word substitutions that represent the specific named error pattern.

Pupil data were preprocessed using the same data cleaning pipeline described for Experiment 1, to remove blinks and noise. Following preprocessing, there were two main styles of analysis. The first was the same as what was used for Experiment 1: baseline-divisive pupil dilation. To simplify the presentation of data, the mean proportional change in pupil dilation was calculated over the window extending from –0.5 to 2.2 s relative to the sentence offset (rather than expressing the full time course for each of the many error types that will be reported). This window includes the growth to peak dilation and the retention interval up to the point where the visual prompt affects pupil size.

The second analysis focused on the impact of context and followed the approach used previously by our lab (Winn, 2016; Winn & Moore, 2018; Winn & Teece, 2020) to calculate the effect of context on pupil dilation. Low-context sentences lack the cues that enable predictions of subsequent words and therefore should demand greater sustained vigilance or extra cognitive control. To express the reduction of effort resulting from high-context semantic cues, we calculated the linear difference between baseline-divisive pupil dilation values (i.e., low-context response minus high-context response) divided by the peak pupil dilation value for low-context sentences. Essentially, this translates to: Given the range of pupil dilation that is elicited without any context, what proportion of that range is reduced by the presence of semantic context? Typically, context results in lower pupil dilation (Winn, 2016), and the degree and timing of that reduction depend on various properties of the stimulus (e.g., speaking rate/style; Borghini & Hazan, 2020) and of the listener (e.g., hearing status: Winn, 2016; native language: Borghini & Hazan, 2020). The reference point of the peak response for low-context stimuli results in a calculation that is self-normalized for differences in language-evoked pupil reactivity across participants. That is, the measurement is expressed as a proportion of the same measurement in the same task in the same listener, elicited by a nominally neutral stimulus type.

For each of the data comparisons that follow, interpretation of the statistical analyses should be done with caution due to the combination of many comparisons as well as the post hoc unbalanced nature of the analysis. Unlike Experiment 1 which had a roughly equal contribution from each participant in each condition, and where specific mistakes were planned and elicited, the second experiment was an unplanned amalgamation of tagged trials from many listeners, without any a priori design. In addition, it cannot be known with certainty that correct responses in Experiment 2 were the result of correct perceptions or if the listener mentally corrected misperceptions.

Experiment 2: Results

Intelligibility. Figure 3 illustrates the diversity in the types of errors that occurred. Phonetic-driven errors were common, with greater prevalence of errors on word-final consonants (consistent with results by Dubno & Levitt, 1981), and relatively fewer errors on vowels. Notably, there were numerous errors that cannot be reasonably described in terms of phonetic misperceptions. For example, the stimulus *The doctor prescribed the drug*, was repeated with *pill* substituted for *drug*; this was classified as a semantic substitution rather than an acoustic-phonetic mistake. In many cases, errors driven by semantic coherence still had some phonetic resemblance with the actual stimulus (e.g., *We heard the ticking of the clock* repeated as *We heard the chicken and the cluck*). The prevalence of errors driven by semantic coherence was unsurprisingly higher in high-context sentences but was not absent in low-context sentences; nearly, one tenth of the errors in low-context sentences were of the semantic variety (typically intrusion of a word earlier in the sentence that was coherent with the final word).

The patterns in intelligibility demonstrate the lack of independence of errors across the sentence, illustrated in Figure 4. First, for both early-occurring and later-occurring words in a sentence, there is a significantly higher likelihood of an error when there was an error elsewhere in the sentence. This pattern was previously reported by Winn and Teece (2020) as having a statistical effect on intelligibility that exceeds the effects of either semantic context or speaking rate. When all preceding words were repeated correctly, sentence-final high-context target words were repeated with 97.8% accuracy, but when there was at least one error before the final word, scores for the final word dropped to 56.5% (Figure 4, right panel). For low-context sentences, the corresponding accuracy rates were 60.6% and 48.3%, respectively. Together, this pattern shows that errors on words early in the sentence are more costly when those words were coherent with the rest of the sentence; mistakes on words are less costly when the rest of the sentence did not depend on correct perception of those words.

The interaction of intelligibility errors across the sentence can also be expressed in the reverse direction; when the final word in high-context sentences was repeated correctly, 88.4% of those responses also had no errors on any of the words leading up to that final word. Conversely, only 10.1% of responses had perfect intelligibility on leadup words when there was an error on the final word (Figure 4, black lines in left panel). This pattern is consistent with either misperceived context leading to a misperceived final word, or a misperceived final word tempting the listener to substitute one of the preceding words to render it coherent with the

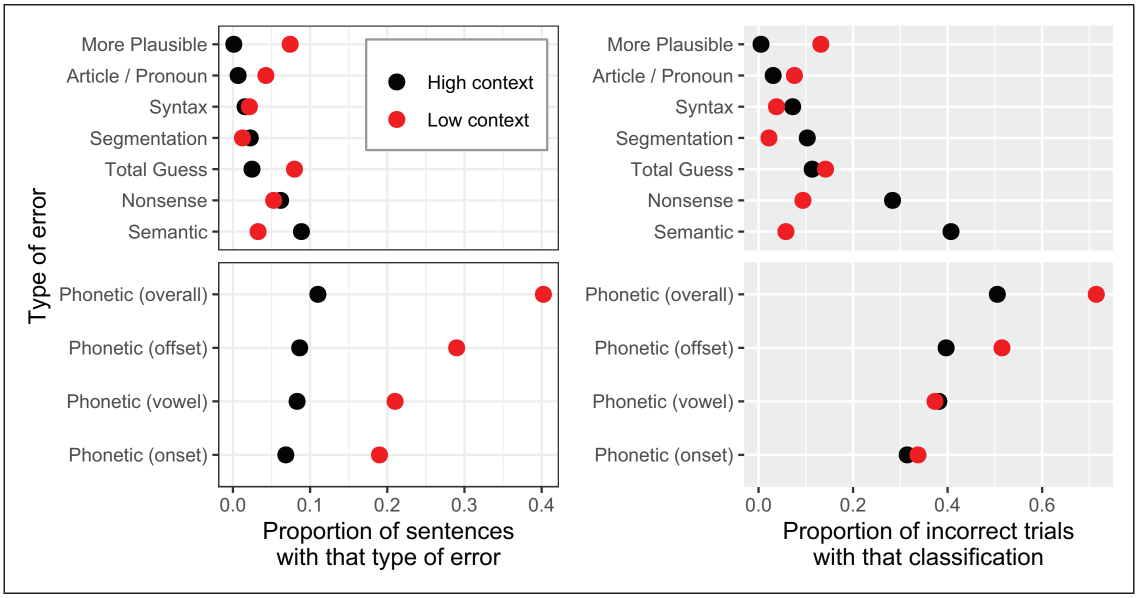

Figure 3. Left panels: proportion of sentences that contained various types of errors, which were not mutually exclusive. Right panels: prevalence of each classification as a proportion of the total number of trials *with errors*.

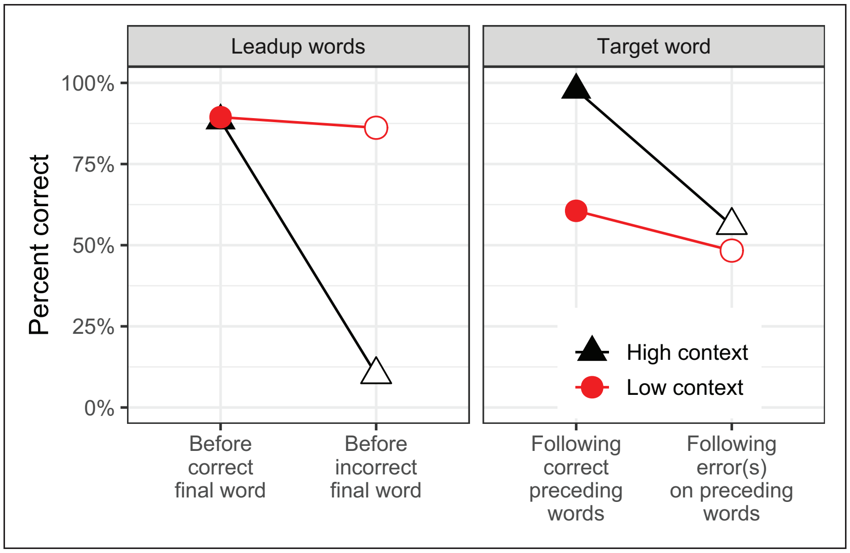

Figure 4. Intelligibility of Two Different Sections of Sentences (Leadup Words: All Words Before the Final Word; Target Word: the Final Word of the Sentence). Filled points indicate that the *other* component was repeated correctly, and open points indicate that the other component contained at least one error.

misperception. Unsurprisingly, the effect of the final word on the success of perceiving preceding words was very small in low-context sentences, where 89.4% of responses contained all correct words preceding a correct target, with only a slight drop to 86.2% correct when the target was incorrect.

Pupillometry. Figure 5 illustrates the results of several statistical models conducted to address specific comparisons in the data. In each panel, proportional change in pupil size is indicated by the *Standard* row, which is the intercept of the model. Every other effect is the *change* to that intercept. An effect whose confidence range that overlaps with zero is likely not a meaningful effect. Effects greater than zero are those that increase pupil size, while effects less than zero are those that decrease pupil size.

The presence of an error generally led to larger pupil size (Figure 5, row 2). However, among trials with an error, the *number* of errors within the sentence did not have a statistical effect on pupil size (row 5), being overpowered by the status of the error being linguistically incoherent (row 4). For example, the stimulus *Jane wants to speak about the chip* was incorrectly repeated as *Gene lost his bag around the gym*; this response contained many errors but was still coherent (i.e., a Type 4 response). Conversely, when the stimulus *The glass had a chip on the rim* was incorrectly repeated as *The glass had a check with the wind,* there were fewer errors but lacked sensible meaning (i.e., Type 5 response). There appears to more effort associated with the incoherent responses, supporting Hypothesis B.

Although the number of total errors did not have a statistical effect, the number of errors *before the final word* did have an effect (Figure 5, row 8), which was counteracted by a negative interaction when the final word was *also* incorrect (row 10). That is, the detrimental effect of an error before the final word was alleviated when the final word was also incorrect, presumably because the two errors were coherent with each other despite being objectively wrong with regard to the actual stimulus. Both of these effects were stronger when tallying the *presence* (rather than the *number*) of errors before the target word, as indicated by rows 12 and 14 (which are respectively stronger than effects in rows 8 and 10, respectively). Together, these effects are inconsistent with an account of effort as a function of the number of errors, and more consistent with an account of effort resulting from listeners attempting to construct coherence in their perceptions. Effort is elevated even when those attempts are successful (Experiment 1), and it appears to be also elevated when the listener fails to give a coherent reconstruction. One of the limitations of this analysis is that the number of errors was calculated in the response, but it cannot be known whether the same number of words were mistaken *in perception*, or whether some words were mentally corrected or mentally overruled to cohere with another part of the sentence.

Context had a predictable effect, with low-context sentences yielding reliably higher pupil responses even when repeated correctly (Figure 5, row 16). An error before the final word (row 17, a reestimation of the effect in row 12) was detrimental in high-context

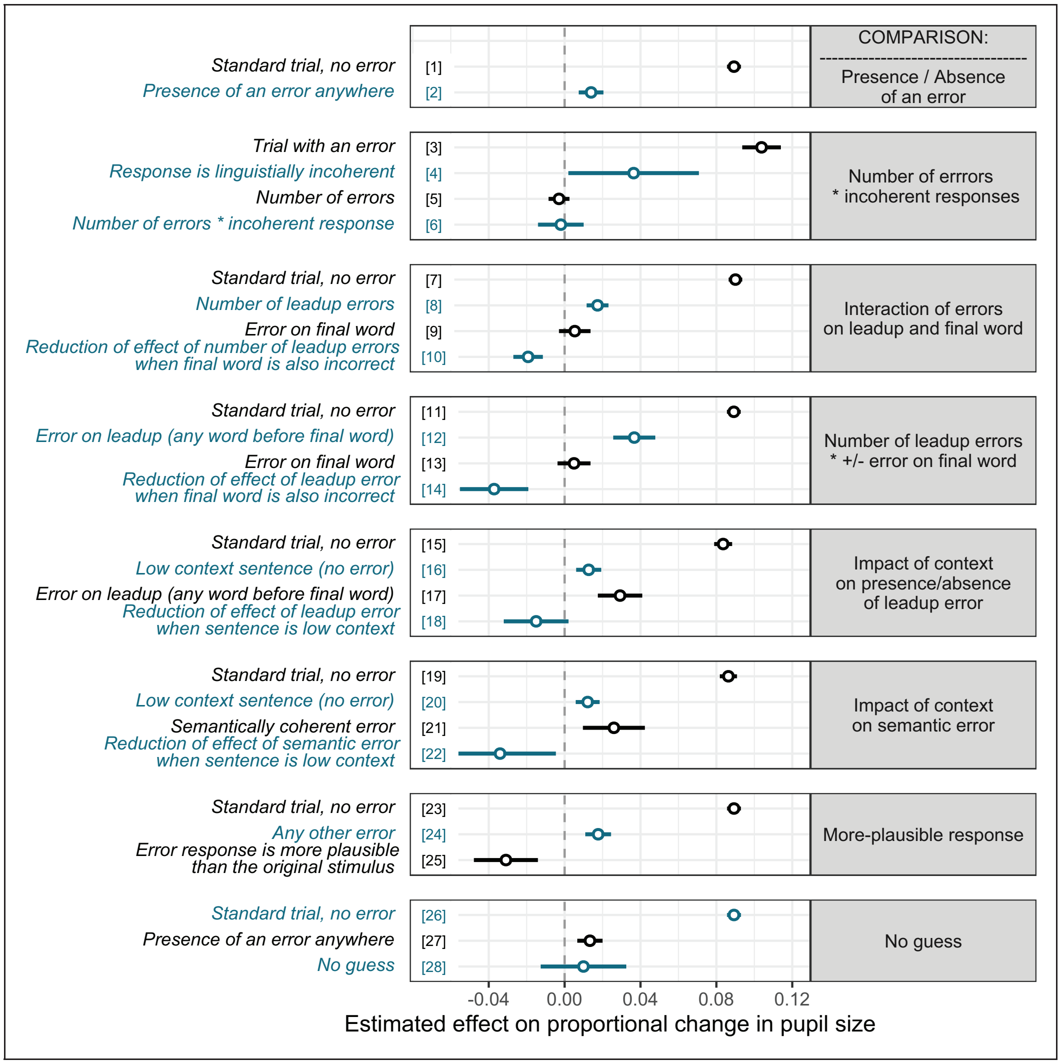

Figure 5. Effect of Various Types and Configurations of Intelligibility Patterns on Pupil Size. Each enclosed panel represents a different self-contained analysis. Standard trials represent the proportional change in pupil size for the trials with no errors (unless specified), and other effects represent change from those default values. Open circles represent point estimates, and line ranges reflect the 95% confident intervals. Colors alternate merely for visibility and do not map to any variables or results.

sentences (high-context being the default level in the statistical model, hence the absence of its designation in the chart row). However, that main effect was counteracted by a negative-direction interaction with low context (row 18), suggesting that the presence of an error before the final word is less costly when the sentence is low context.

Erroneous word substitutions were often semantically coherent with other words in the sentence. For example, the high-context stimulus *Kill the bugs with this spray* was repeated as *Fill the bottle with this spray*, and the stimulus *The bride wore a white gown* was repeated as *The bride wore a white veil*. In such cases, there was an increase in pupil size for high-context sentences (Figure 5, row 21). However, a semantic error is less costly when the stimulus was originally a low-context sentence, as indicated by the negative interaction of the error with low context (row 22). Averaging across both context types, semantically coherent word substitutions (response Types 3 and 4) tend to elicit only a moderate amount of additional effort compared to other errors, likely because they tend to retain linguistic coherence. Low-context sentences were sometimes rendered *more* coherent (i.e., more plausible) with the addition of errors early in the sentence. For example, *He's glad you called about the jar* repeated as *His dad is thrilled about the job* or *He's glad that you could open the jar*. When responses were more plausible than the original stimulus, the effect on pupil size was negative, relative to correct responses (row 25), implying that the tendency to construct more-likely sequences of words could be an automatic process that demands less effort than intentional analytic processing of sentences with less-predictable sequences. In the case of the sentences being rendered more plausible, it might not have been a total reconstruction but rather an automatic process that associated some key words with other coherent-related concepts. Hypothesis B (predicting that semantically coherent errors would be less effortful) was thus partly supported.

On occasion, a participant would simply not even offer a guess. Whereas the presence of an error in an actual verbal response yielded a reliable increase in pupil dilation compared to correct responses (row 27), the absence of a guess yielded changes in pupil size that were inconsistent and not statistically different from correct responses (row 28). This might be because some nonguesses result from complete confusion, while others might result from temporary lack of attention (cf. Breeden et al., 2017).

A separate statistical model included terms that were not included in any of the planned comparisons, including errors on syntax, articles/pronouns, segmentation errors, semantically coherent errors, unrelated guesses, and phonetically-similar errors. None of these types of errors resulted in any statistical effect (all $|t|$ less than 0.8; all $p > .4$), implying that their impact is best expressed as an interaction with the rest of the sentence, either through linguistic coherence, or through a relationship with other errors in the sentence. We refrained from including the full matrix of multiway interactions in this model, to retain its simplicity and stable convergence.

Data were analyzed to further illustrate the main patterns described earlier in the statistical results. Figure 6 demonstrates clearly that there is no consistent trend of pupil dilation across the number of errors, but there is a consistent trend of greater dilation when the response is linguistically incoherent, consistent with Hypotheses B and C. This suggests that Type 5 errors are reliably more effortful than Type 3 or 4 errors. Specific to Hypothesis C, a single error that results in an incoherent response appears to elicit greater effort than more errors within a coherent response.

As laid out in the Introduction, response Type 6 corresponds to the situation where one error appears to result from the listener's accommodation of another error elsewhere in the sentence. Consistent with the previous examples, Figure 7 displays the trend that when there are errors earlier in the sentence, an *additional* error on the sentence-final word tends to be associated with *reduced* pupil dilation, regardless of the number of leadup-word errors. Taken together with the prevalence of errors that retain some semantic coherence within a sentence, the provisional interpretation of this trend is

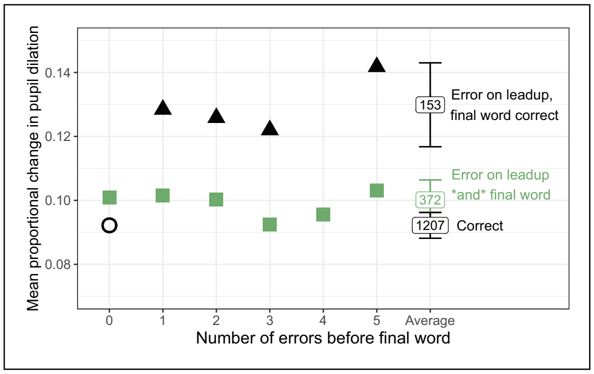

Figure 7. Mean Pupil Dilation Associated With Various Numbers of Errors Before the Final Word, Split by Whether Those Errors Were Followed by an Error on the Final Word. The enclosed number indicates the number of trials included in each average. Error bars represent ±2 standard errors.

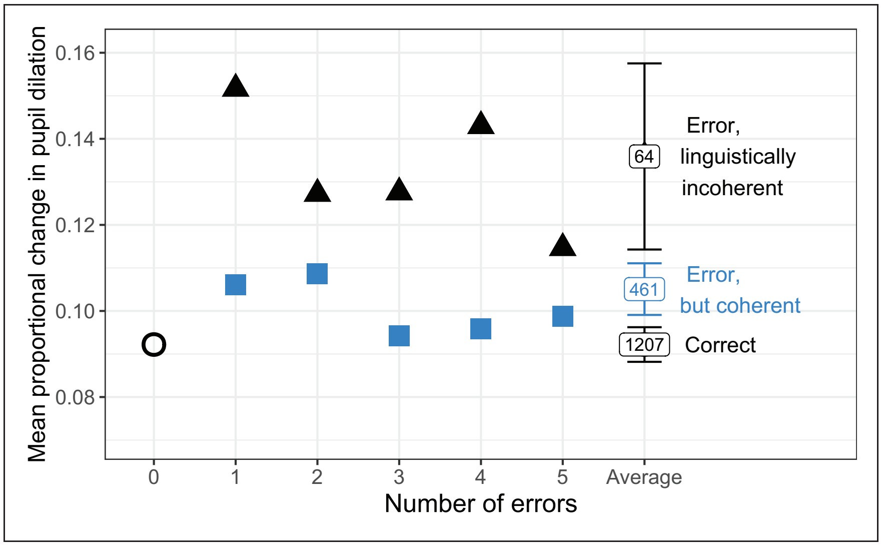

Figure 6. Mean Pupil Dilation Associated With Various Numbers of Errors, Split by Whether Those Errors Results in a Coherent or Incoherent Sentence. The enclosed number indicates the number of trials included in each average. Error bars represent ±2 standard errors.

that additional errors can decrease listening effort if those errors are coherent with other (misperceived) parts of the sentence. Hypothesis C is thus partially supported.

Breakdown of Errors Earlier and Later in the Sentence, Interacting With Context. To address Hypothesis D, Figure 8 illustrates the changes in pupil dilation elicited by errors at earlier and later parts of high- and low-context sentences. Section A aggregates over each type of sentence context. Data from NH listeners in Panel 1 are data collected in the study by Winn (2016) and are illustrated to represent the ideal case of sentence processing with no auditory dysfunction; pupil dilation is reduced when the stimuli are high-context sentences. The complete data set for CI listeners in Figure 8 (Panel 2) shows less separation between high- and low-context responses, but this separation is slightly larger for trials in which all of the words were repeated correctly (Panel 3). Trials in which the target word was the only word reported incorrectly yielded data for which there was essentially no impact of context on pupil dilation (Panel 4). Conversely, errors among the words earlier in the sentence led to a much larger change in pupil dilation. The pattern of smaller relative pupil dilation for high- versus low-context sentences was not only nullified, but *reversed direction* when there was an error among the words preceding the target word (Panel 5). Surprisingly, the magnitude of this reversal (i.e., the cost of the leadup word mistake) is *diminished* when the target word is *also* incorrect (Panel 6) but grows even larger when the target word is correct (Panel 7). However, one limitation here is that the presence of perceptual errors cannot be ruled out in the case of correct responses, so the impact of each error type is possibly underestimated in this analysis.

Figure 8 Section B has the same *x*-axis (timeline) as Section A above it, and directly represents the separation between curves in Section A, proportional to the low-context reference (described earlier in the analysis section). The peak value for the low-context stimuli was nearly identical across all data subsets, validating its status as a reliable neutral comparison and simplifying the comparison across response patterns. In this figure, the "zero" represents the situation in which high-context sentences elicit the same pattern of pupil dilation as low-context sentences. If high-context sentences elicit reduced dilation (as one would expect), the data fall below the zero line; this pattern is observed for the NH listener data set. In previous publications, we have termed these relative reductions in pupil dilation "effort release" because it appears that successful perception of semantic context offers release from effort during speech processing (similar to how spatial separation between target and masker offers release from masking). The current analysis reveals a more complicated picture; there are circumstances where misperception of context results in *increased* pupil dilation (i.e., the data rise above zero) for high-context stimuli. Specifically, an

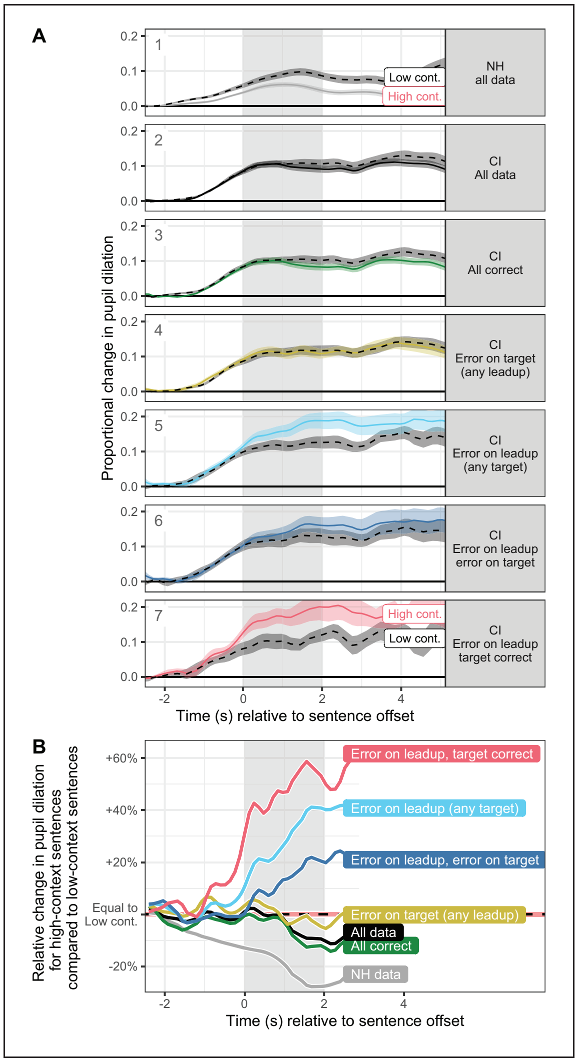

Figure 8. Panel A: Changes in pupil dilation for high-context (solid colored lines) and low-context (dashed black lines) sentences, with various data subsets split into different panels. Error ribbons represent ±1 standard error. Target refers to the sentence-final word, and "leadup" refers to any of the words preceding the target. "Any {target/leadup}" means data were averaged over both correct and incorrect items. Panel B: difference between pupil dilation responses for high- and low-context sentences displayed in Panel A, expressed as proportional change of the high-context response relative to the peak dilation in the low-context response. Data below 0 represent reduction of pupil dilation for high-context relative to low-context stimuli, and data above 0 represent increased pupil dilation for high-context stimuli. In both panels, the gray shaded region indicated the 2 s of silence separating the end of the stimulus to the onset of the response prompt.
NH = normal hearing; CI = cochlear implant.

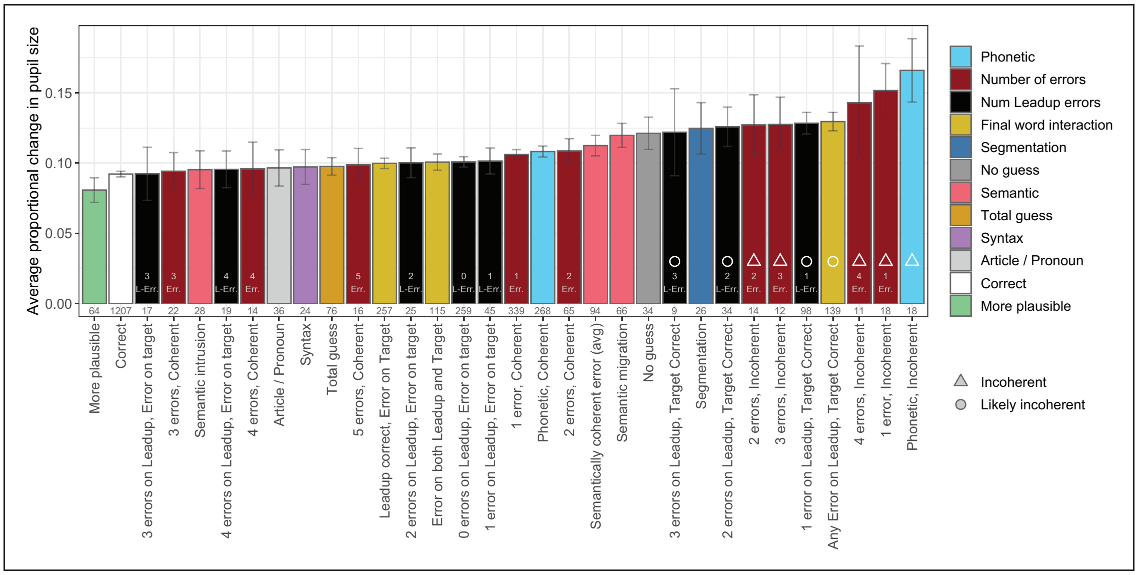

Figure 9. Mean Proportional Change in Pupil Dilation Associated With a Variety of Response/Error Types. Encircled numbers represent number of trials contributing to each data point. Open triangles mark error types where responses are linguistically incoherent (i.e., nonsense utterances), and open circles indicate response categories that were likely to contain some incoherent responses but not guaranteed to all be linguistically incoherent (e.g., high-context sentences with errors on early-occurring words, but correct final words). Error bars reflect ±1 standard error.

error early in the sentence results in high-context sentences being more effortful to process, likely because the error was a word that was crucial to the coherence of the sentence, and is rendered incoherent with the final word.

Overall Comparison of Many Types of Responses. Experiment 2 contained a diverse set of responses and error types that could be used to explore the relative difficulty of mistakes that were not prospectively built into the experiment. These errors were subject to comparison, with the caveat that they were variable in number, not evenly represented across participants, and not explicitly controlled as they were in Experiment 1. Figure 9 displays mean pupil dilation for each response type during the window spanning from –0.5 to 2.2 s relative to sentence offset for 32 different non-mutually-exclusive response patterns. These data were subject to the same preprocessing as each of the other data points elsewhere in this study. The number of trials contributing to each data point is indicated by the small numbers below each bar. The responses that trend toward incoherence dominate the upper end of the range of pupil dilations, with no clear pattern for *number* of errors.

Phonetic errors that still produce a coherent/sensible sentence (Type 3 response, e.g., *Our seats were in the second row* repeated as *Our seats were in the center row*) appear to elicit pupil dilation that is barely different than that from nonphonetic errors (Type 4), and only slightly more than correct responses. However, when phonetic errors result in nonsense utterances (Type 5, e.g., *We played a game of cat and mouse* repeated as *We played a game as cough and mouse*), there is a much larger increase in pupil dilation. Semantically coherent errors were further distinguished as intrusion of contextual words into low-context sentences (which was relatively less effortful) or migration of semantic context within a high-context sentence (which was relatively more effortful).

Unlike the errors that resulted in incoherent responses, incorrect reporting of articles and pronouns was associated with minimal changes in pupil dilation. Mild effects were also observed for syntactic errors as well as the intrusion of semantically coherent words in low-context sentences. However, errors on word segmentation (e.g., *Unlock the door* … as *I locked the door* … , *The dealer shuffled* … as *The dealership* …), result in larger dilations, for reasons that are not immediately clear. This type of error was previously tracked by Perry and Kwon (2015), who found a relationship between segmentation strategy and the ability to hear speech in noise.

Experiment 2: Discussion

There are four main themes that emerge from the second experiment, which help to clarify the relationship between listening effort and different types of speech perception mistakes. First, effort is not related to the number of errors in a response but is related to whether the listener needs to engage in corrective action to produce a linguistically coherent response (Figures 6 and 9). Second, errors earlier in a sentence are more costly than errors late in a sentence (Figure 8). Third, misperceiving context is more costly than not having any context (Figure 8). Fourth, the bias toward constructing responses that are plausible and coherent can override auditory perception—producing errors that cannot be predicted by phonetic error patterns and will sometimes result in *more* errors that are *less* effortful to process (Figures 8 and 9). Unsurprisingly, it appears as though listeners are biased toward hearing sensible and ordinary sentences that demand less effort, and they will impose semantic coherence even when it is lacking in the original signal. Collectively, these observations help to show that effort and intelligibility scores are separable concepts.

One of the largest effects observed in Experiment 2 was that errors are more effortful when they occur earlier in a sentence (Figure 8). This trend is consistent with several foundational ideas in psycholinguistics: Sentence processing is governed by active predictive processing (Altmann & Kamide, 1999; Federmeier, 2007; Kamide, 2008), and *garden-path* sentence parsing that requires backtracking and rebuilding is especially cumbersome (Ferreira & Clifton, 1986; Ferreira et al., 2001). The increased effort likely reflects the cost of overcoming the initial automatic predictions generated from the misperceived words (Shenhav et al., 2017). There are entire paradigms of research based on the concept of building and then intentionally violating a listener's expectations to demonstrate the influence of language on predictive and restorative processing (Kutas & Hillyard, 1984), including studies that track the resolution of sentence structure in real time (MacGregor et al., 2019). The study of speech perception by individuals with hearing impairment offers a special opportunity to observe the consequences of these predictions and violations while still using typical sentences that have the desirable quality of maintaining the listener's normal expectation of coherent sensible stimuli. This result highlights the fact that a linear sum of errors (traditional intelligibility scoring) overlooks the palpable differences in the consequences of different errors.

Contextual information within a sentence has a strong influence on effort and interacts with the presence of errors. When words allow successful prediction of later words, this lowers the utility of continued attention (cf. Shenhav et al., 2013; Tversky & Kahneman, 1974) so that persistent effort is suboptimal and thus withheld (Griffiths et al., 2015). Thus, rather than exerting *extra* effort to process the language, the semantic cues are used as a simple heuristic driving downregulation of subsequent attention (Gigerenzer, 2008). Consistent with this, contextual information rapidly reduces effort during sentence processing if the signal is heard cleanly but less rapidly when the signal is degraded (Winn, 2016). Context can be detrimental if it leads to the *wrong* predictions, such as when an informative word is misperceived in a way that is no longer compatible with the rest of the words in the sentence. Some of this impact can be seen even in the intelligibility scores: When context was misperceived, intelligibility for high-context target words was just as poor as for low-context target words (Figure 4). This observation is consistent with the analysis and modeling by Marrufo-Pérez et al. (2019), who showed systematically worse performance on words in Spanish-Harvard sentences when those words were immediately preceded by an error on the previous word. Their study suggests that individual trial analysis yields rich data that might be lost when aggregating intelligibility over an entire experimental condition. In agreement with the conclusions by Marrufo-Pérez et al., intelligibility of both early- and late-occurring words in sentences in the current study were heavily influenced by the presence of an error elsewhere in the sentence (Figure 4). As they note, *although speech predictability can facilitate sentence recognition, it can also result in declines in word recognition as the sentence unfolds because of inaccuracies in prediction* (Marrufo-Pérez et al., 2019, p. 1).

Perhaps the most surprising result of Experiment 2 was that following an error early in the sentence, an additional error at the end of the sentence resulted in *reduced* pupil dilation, while the *correct* repetition of the final word resulted in greater pupil dilation (Figures 7 and 8). A likely explanation for this is that a single error results in the earlier and later parts of the sentence to be in conflict, while a second error could render the sentence coherent, even if it is incorrect. For example, in stimulus *We were considering the cheers* repeated as *We were sitting in the chairs* is rendered more plausible because of the second error (chairs instead of cheers). Although *They were sitting in the cheers* would only have one error, it would likely elicit relatively more effort because the ambiguity was not resolved. The increased pupil dilation in the case of single errors could therefore indicate failure to resolve sensible meaning rather than the mere status of the perception as incorrect.

There are at least two important factors that remain unexamined in Experiment 2. First, it cannot be determined whether the stimuli that were repeated accurately were initially perceived correctly. This means that the patterns of mistakes organized in the analyses and figures represent mistakes that were uncorrected and overlook some mistakes that were corrected by the listener before giving a response. In addition, part of the elevated pupil response in the incoherent sentences could have resulted from the added difficulty of *producing* an incoherent response. Although those elevated responses appeared well in advance of the prompt to give a verbal response, it cannot be determined from these data when speech motor planning began.

General Discussion

This study was carried out with the goal of understanding changes in listening effort grounded not in correlations with intelligibility scores, but instead based on an explanatory account of what types of perceptions and misperceptions arise, and what kind of cognitive load is demanded or relieved by each type. In numerous distinct ways, the current study shows that the number of errors in a response is not a sufficient explanation of changes in listening effort. Correct responses could be effortful if they involve mental correction of misperceived words in a sentence (Experiment 1, Figure 2). Misperceived words are more costly when they occur earlier in a sentence (Experiment 2, Figure 8), and when there is an error, the addition of another error can reliably reduce effort (Experiment 2, Figure 7). Overall, elevated effort is better explained by the need to resolve linguistic ambiguity rather than a greater number of errors. The elevated effort of mentally correcting a misperception is brief when the ambiguity is successfully resolved (Experiment 1) but is prolonged and further elevated when the listener has not successfully resolved the ambiguity to arrive at a sensible or meaningful sentence (Experiment 2, Figures 6 and 9). The importance of linguistic coherence is further supported by the notion that errors that are more coherent than the actual stimulus elicit smaller pupil dilation than genuine correct responses (Experiment 2, Figure 9). Although percent-correct intelligibility is a much simpler calculation to make, and will *tend* to scale with effort (Wendt et al., 2018; Zekveld et al., 2010), the reason for this correlation is likely because listening conditions lead to a greater number of intelligibility errors also gave listeners a greater number of opportunities to exert effort by attempting to resolve incoherent misperceptions (rather than the mere fact of the presence of errors). Although linguistic processing is not the only source of effort, the contributions of linguistic processing explain more than a linear tally of the number of errors.

Rönnberg et al. (2013, 2019), suggested that explicit cognitive resources are selectively activated to reconstruct fragmented phrases and to support inference-making. The empirical data in the current study directly support that idea, both in terms of reconstructing phrases with missing pieces (Experiment 1) and making inferences constrained by properties of the language (Experiment 2). Normal language processing should be quick and efficient when the signal is clean and well-formed, and listeners with NH indeed show reduced effort when listening to sentences with linguistic coherence (Borghini & Hazan, 2020; Winn, 2016). But when the signal is degraded either systematically via background noise (Zekveld et al., 2010), spectral degradation (Winn et al., 2015), divided attention (Koelewijn et al., 2015, 2017), or selective word masking (Experiment 1 in the current study), momentary engagement of effort is observed in the data. This framework is consistent with previous studies that characterize pupil dilation as reflecting active decision-making processes (Cavanaugh et al., 2014). Engagement of those processes appears to be malleable, as Hsu and Novick (2016) showed that priming listeners with a task that specifically stimulates active cognitive control (a Stroop task) can facilitate earlier recovery of correct perceptions following initial incorrect perceptions. In both Experiment 1 (explicitly) and Experiment 2 (implicitly), we suspect that the process of generating meaning out of an incomplete perception was the effortful part of the experience (rather than the simple fact that an error was made). Because the largest effects in Experiment 2 resulted from incoherent responses, we posit that the mental correction process was not absent in those responses but rather remained active until the listener simply decided that they were not able to add any more coherence to the perceived sentence. It is not possible to be certain of this idea, though it is consistent with the data collected by Bradshaw (1968) who tracked the timing of solutions to math problems in a study that also illustrated sustained elevated responses for problems that participants ultimately failed to solve.

Deconstructing the Reductionist View of Speech

Results of this study suggest that a thorough understanding of speech perception cannot be achieved by an atomic/reductionist model based on perception of independent units such as consonant and vowel features, or single words. Even though detailed analyses show systematic patterns in the perception of phonetic constituents of isolated syllables (e.g., Dubno & Levitt, 1981; Miller et al., 2017; Toscano & Allen, 2014), the data presented here suggests that other factors contribute meaningfully and can override the perception of phonemes. Many of the patterns in the current study could not emerge without multiword utterances that could potentially have internal coherence or incoherence. There was a substantial number of errors that were consistent with the semantics of a sentence but not the phonetics of the misperceived word (Figure 3). Among the high-context sentence trials with errors, 81% contained at least one error other than a phonetic mistake. The corresponding number for low-context sentences was 56%, indicating that sentences that were mostly devoid of semantic processing still contained a very large proportion of mistakes that would not be predictable just from phonetic misperceptions. These patterns are consistent with previous observations that word errors are likely not entirely predictable from perceptions of phonemes in isolation (Boothroyd & Nittrouer, 1988), especially for listeners with CIs (Gianakas & Winn, 2019). Kaandorp et al. (2017) found that predictors of sentence recognition were statistically unrelated to predictors for digit recognition and single-word perception. In addition to the disconnect between different kinds of utterances, the examples shared in the current article suggest that a participant's auditory perception cannot be linearly decoded based on verbal responses—both mistakes and correct repetitions can reflect cognitive transformations of auditory percepts. Furthermore, those transformations appear to underlie some of the larger effects on listening effort observed in this study.

The Implications of This Study for Speech Perception Testing in General

Emergent linguistic properties in sentence-length utterances are not merely curiosities or noise in the data—they appear to carry the load of determining how much effort is exerted. Therefore, speech stimuli designed to avoid the linguistic variety of natural speech (e.g., monosyllabic words, digits) or to avoid the unconstrained variety of open-ended responses (e.g., the Coordinate Response Measure by Bolia et al., 2000, and Oldenburg matrix sentence test by Wagener et al., 1999) are not simplifying the search for listening effort—they are likely addressing different questions unrelated to the effort of language processing. Such tests can be useful as probes of the auditory system since they show reliable reductions in performance with background noise and can elicit effort of signal detection and digit recognition (Mackersie & Cones, 2011), as well as recognizing the affective response to noise (Francis et al., 2016; Love et al., 2019). However, closed-set responses by definition exclude the recognition of the types of errors that elicit meaningfully different amounts of effort. Closed-set response tests that enforce linguistic coherence in the response or are linguistically inert (e.g., digits) are destined to overlook the type of linguistic processing highlighted in this study, because the responses are constrained to all be coherent (i.e., response Types 2, 3, or 4), regardless of the listener's perception. A problematic feature of closed-set tests is that they do not allow Type 5 or Type 6 errors, which are the responses that carried much of the load in explaining effort in the current study. Consistent with the main idea of the "selective gain" mechanism described by Kerlin et al. (2010), we contend that speech perception stimuli should sufficiently engage cognitive decision-making processes and invite the listener to construct meaningful coherence if effort is the target measurement.

If effort scales closely with perceived linguistic coherence, then there is extra value in using sentence-length stimuli that give the listener an opportunity to parse that coherence, compared to single words or sequences of unrelated syllables. Functional imaging studies show brain regions that selectively respond to syntactic organization and semantic integration (Rogalsky & Hickok, 2009). Predicting and reconciling sentence content on a full-utterance level (rather than the individual word level) understandably has a substantial impact on effort. For example, spectral degradation heavily impacts the effort of perceiving sentences (Winn, 2016; Winn et al., 2015) but has no statistical effect of recognition of individual syllables (McCloy et al., 2017). McCloy et al. describe the results of the Winn et al. (2015) study as follows:

> One might say that signal degradation itself was not the proximal cause of pupil dilation in those sentence comprehension experiments; rather, it was the additional cogitation or effort needed to construct a coherent linguistic meaning from degraded speech that led to the pupillary responses they observed.

Echoing the earlier statement about the value of mistakes in perceptual tests, it is worth considering how mistakes arise. The approach used in Experiment 1 does not contain stimulus degradation that will *likely* lead to mistakes (e.g., continuous background noise, spectral degradation) but rather is designed to *inevitably* result in specific mistakes so that the ensuing mental correction process can be experimentally controlled. In other words, Experiment 1 simulated mistakes directly rather than imposing conditions that lead to mistakes outside the experimenter's control. Experiment 2 was less controlled; the influence of language processing means that the responses cannot be taken to be a perfect representation of the perception, so the analysis for Experiment 2 should be considered exploratory.

Despite the emphasis on the diversity of intelligibility-effort patterns in the current study, the current experiments are still an incomplete representation of communication in the real world. Conversation involves connecting meaning within and across utterances, revising perceptions as new information arrives, and incorporating background knowledge, all while preparing a creative or informative response rather than verbatim repetition. Gatehouse (1998) advocated for greater realism of linguistic properties in speech perception materials. This view was endorsed by Best et al. (2016) and Beechey et al. (2019), who made a compelling case for striving toward more realistic communication scenarios other than verbatim repetition (including informative dysfluencies and pragmatics). The effort of communication is not necessarily related to every word that is spoken but is also a reflection of the listener's personal experience and comfort with uncertainty (Francis & Love, 2019). These concepts are all challenging to implement experimentally but are a necessary part of the mission of understanding and ultimately alleviating listening effort.

The Evolving View of Listening Effort

One of the more convincing illustrations of the value of measuring listening effort is when effort is different despite equivalent intelligibility scores (e.g., Koelewijn et al., 2012). The current study highlights how a priori requirement of matching intelligibility scores is problematic in principle. Put simply, a listener's response is not a reliable readout of their initial perception, since the linguistic constraints that govern their perception will transform their verbal response as well, guiding it toward a coherent form like an attractor in a dynamical system. To borrow terminology from signal detection theory, the intelligibility score as measured by the experimenter is an accumulation of some genuinely correct perceptions (hits), some incorrect perceptions that were corrected by context or intuition (misses), some genuinely incorrect responses (correct rejections), and some correctly perceived words that were repeated incorrectly because the listener transformed them to be coherent with something elsewhere in the sentence (false alarms). There is no assurance that the number of false-alarm correct perceptions will equal the number of misses and also no assurance that they play equivalent roles, even if equal in number. The intelligibility score thus offers no guarantee as to how much of the signal was accessible, versus how much was reconstructed by the listener. Equating intelligibility is further complicated by the fact that different ways of changing intelligibility present different challenges to the listener, which can lead to different kinds of effort (Francis et al., 2016; Strauss & Francis, 2017). Decreasing intelligibility by adding noise could lead to a qualitatively different experience than listening to speech whose intelligibility is equated by bandwidth reduction, reverberation, spectral distortion, accentedness, or increased speaking rate. For example, if the experimenter must impose a very poor signal-to-noise ratio to reduce digit-recognition performance down to 50%, there will be changes in effort (Mackersie & Cones, 2011) but that effort might reflect the nuisance of noise and source segregation (Love et al., 2019).

Modern frameworks of listening effort (cf. McGarrigle et al., 2014, Pichora-Fuller et al., 2016) are complex and multidimensional; the linguistic processing described in the current study is only a fraction of the larger problem. In a large-scale study involving seven different measures of effort, five tests of cognitive function, and two personality measures, Strand et al. (2018) showed weak or absent correlations between different measures, suggesting that they are tapping into different phenomena. A majority of their cognitive and personality predictors showed stronger relationships to listening effort when the speech task was more difficult. Alhanbali et al. (2019) provided further evidence of the multidimensionality of effort, suggesting that different objective measurement techniques (e.g., EEG, skin conductance, pupil dilation) could tap into independent components of effort and/or fatigue (see also McGarrigle et al., 2017). It is possible that the effort measured in the current study reflects only the cognitive load associated with decision-making/ambiguity resolution, but not the other components. Some deeper explorations of listening effort expand into cognitive factors such as verbal working memory and attentional-based performance control (cf. Peelle, 2018). Differences in the effort of *detection* and the effort of *processing* were previously underscored by Beatty (1982), who showed that a detection task elicited substantially smaller pupil dilations compared to tasks that involved active cognitive processes and decision-making. Some listeners can deploy their effort more efficiently than others, as shown by pupillometry tasks by van der Wel and van Steenbergen (2018) and through a variety of behavioral tasks by Strand et al. (2018). In everyday communication, it is reasonable to suspect that individuals with hearing loss can momentarily increase engagement for important situations but would need to be economical with that effort to avoid debilitating fatigue (cf. Eckert et al., 2016; Winn et al., 2018). Related to this, individuals with hearing loss interviewed by Hughes et al. (2018) alluded to the *efficacy* of effort (not necessarily the *amount* of effort) as a strong factor in their willingness to engage in conversations.

Conclusions

Intelligibility scores are not a sufficient explanation of listening effort, and in fact, these two concepts can be doubly dissociated. In sentence intelligibility tests, correct responses can result from an effortful process of mentally correcting misperceived words to produce meaning, and multiple errors can be less effortful than a single error if that single error results in a linguistically incoherent perception. Furthermore, sentences with an error can be rendered less effortful with the addition of more errors. There is more impact on effort resulting from an error in a word early in a sentence compared to an error later in a sentence, suggesting a linear tally of errors is not an adequate model of effort. Errors are more costly when they result in incoherence between earlier and later parts of the sentence, whereas errors that do not result in incoherence (such as mistaking the final word in a low-context sentence) do not elicit much change in effort. Although more intelligibility errors will likely elicit greater effort on account of the larger number of opportunities for cognitive repair, the correlation between intelligibility and effort does not provide the explanatory mechanism that emerges when framing effort in terms of the listener's need for decision-making, resolving ambiguity in sentence parsing, correcting mistakes, and reconciling semantic incompatibilities.

Responses in speech perception tasks often appear to be unrelated to the acoustics of the speech, and instead reflect the influence of surrounding semantic context and basic knowledge about likelihood of certain words or expressions. Future investigations of speech perception and listening effort can therefore be useful probes for auditory processing and yet remain incomplete if they do not specifically account for these linguistic/cognitive influences, which are meaningful rather than superfluous. Language-related errors occur with ample frequency and have a large effect on cognitive load. Rather than dismissing that fact as an inconvenience, it can be embraced and explicitly studied, toward a more complete understanding of speech perception and listening effort.

Acknowledgments

The University of Minnesota stands on Miní Sóta Makhóčhe, the homelands of the Dakhóta Oyáte. Data collection was assisted by Emily Hugo, Paula Rodriguez, Lindsay Williams, Hannah Matthys, and Siuho Gong. Valuable input to this project was given by our laboratory participants, as well as Peggy Nelson, Timothy Beechey, Erin O'Neill, Steven Gianakas, and Viral Tejani. Deidentified data are available upon request. Stimulus design for Experiment 1 was assisted by our late colleague Akira Omaki.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by NIH-NIDCD R01 DC017114 (Winn).

ORCID iD

Matthew B. Winn https://orcid.org/0000-0002-4237-7872

References

Alhanbali, A., Dawes, P., Lloyd, S., & Munro, K. (2018). Hearing handicap and speech recognition correlate with self-reported listening effort and fatigue. *Ear and Hearing*, *39*, 470–474. https://doi.org/10.1097/AUD.0000000000000515

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

Altmann, G., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/s0010-0277(99)00059-1

Altmann, G., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*, 583–609. https://doi.org/10.1111/j.1551-6709.2009.01022.x

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N. Z., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioural experiment builder. *Behavior Research Methods*, *52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

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

Ayasse, N., & Wingfield, A. (2018). A tipping point in listening effort: Effects of linguistic complexity and age-related hearing loss on sentence comprehension. *Trends in Hearing*, *22*, 1–14. https://doi.org/10.1177/2331216518790907

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

Beechey, T., Buchholz, J., Keidser, G. (2019). Eliciting naturalistic conversations: A method for assessing communication ability, subjective experience, and the impacts of noise and hearing impairment. *Journal of Speech, Language, and Hearing Research*, *62*(2), 470–484. https://doi.org/10.1044/2018_JSLHR-H-18-0107

Best, V., Streeter, T., Roverud, E., Mason, C., & Kidd, G. (2016). A flexible question-and-answer task for measuring speech understanding. *Trends in Hearing*, *20*, 1–8. https://doi.org/10.1177/2331216516678706

Bhargava, P., Gaudrain, E., & Başkent, D. (2014). Top-down restoration of speech in cochlear-implant users. *Hearing Research*, *309*, 113–123. https://doi.org/10.1016/j.heares.2013.12.003

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 & Hearing, 40, 1220–1232.

Bilger, R., & Wang, M. (1976). Consonant confusions in patients with sensorineural hearing loss. *Journal of Speech and Hearing Research*, *19*, 718–740. https://doi.org/10.1044/jshr.1904.718

Bilger, R., Nuetzel, J., Rabinowitz, W., Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *J Speech Hear Res*, 27, 32–48.

Block, C., & Baldwin, C. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research*

Methods, *42*, 665–670. https://doi.org/10.3758/BRM.42.3.665

Bolia, R., Nelson, W., Ericson, M., & Simpson, B. (2000). A speech corpus for multitalker communications research. *Journal of the Acoustical Society of America*, *107*(2), 1065–1066. https://doi-org.ezp3.lib.umn.edu/10.1121/1.428288

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *Journal of the Acoustical Society of America*, *84*, 101–114. https://doi.org/10.1121/1.396976

Borghini, G., & Hazan, V. (2020). Effects of acoustic and semantic cues on listening effort during native and nonnative speech perception. *Journal of the Acoustical Society of America*, *147*(6), 1783–3794. https://doi.org/10.1121/10.0001126

Bradshaw, J. (1968). Pupil size and problem solving. *The Quarterly Journal of Experimental Psychology*, 20, 116–122.

Breeden, A., Siegle, G., Norr, M., Gordon, E., & Vaidya, C. (2017). Coupling between spontaneous pupillary fluctuations and brain activity relates to inattentiveness. *European Journal of Neuroscience*, *45*(2), 260–266. https://doi.org/10.1111/ejn.13424

Cavanaugh, J., Wiecki, T., Kochar, A., & Frank, M. (2014). Eye tracking and pupillometry are indicators of dissociable latent decision processes. *Journal of Experimental Psychology General*, *143*(4), 1476–1488. https://doi.org/10.1037/a0035813

Danermark, B., & Gellerstedt, L. (2004). Psychosocial work environment, hearing impairment and health. *International Journal of Audiology*, *43*(7), 383–389. https://doi-org.ezp1.lib.umn.edu/10.1080/14992020400050049

Demberg, V., & Sayeed, A. (2016). The frequency of rapid pupil dilations as a measure of linguistic processing difficulty. *PLoS One*, *11*(1), e0146194. https://doi.org/10.1371/journal.pone.0146194

Dubno, J., & Levitt, H. (1981). Predicting consonant confusions from acoustic analysis. *Journal of the Acoustical Society of America*, *69*(1), 249–261. https://doi.org/10.1121/1.385345

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.ezp3.lib.umn.edu/10.1044/jshr.2501.141

Eckert, M., Teubner-Rhodes, S., & Vaden, K. (2016). Is listening in noise worth it? The neurobiology of speech recognition in challenging listening conditions. *Ear and Hearing*, *37*(Suppl 1), 101S–110S. http://dx.doi.org/10.1097/AUD.0000000000000300

Federmeier, K. (2007). Thinking ahead: The role and roots of prediction in language comprehension. *Psychophysiology*, *44*(4), 491–505. https://doi.org/10.1111/j.1469-8986.2007.00531.x

Ferreira, F., Christianson, K., Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, *30*, 3–20. https://doi.org/10.1023/a:1005290706460

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*(3), 348–368. https://doi.org/10.1016/0749-596X(86)90006-9

Francis, A., Tigchelaar, L., Zhang, R., Zekveld, A. (2018). Effects of second language proficiency and linguistic uncertainty on recognition of speech in native and nonnative competing speech. *J Speech Lang Hear Res*, epub, 1–16.

Francis, A., & Love, J. (2019). Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdisciplinary Reviews. Cognitive Science*, *11*(1), e1514. https://doi.org/10.1002/wcs.1514

Francis, A., MacPherson, M., Chandrasekeran, B., & Alvar, A. (2016). Autonomic nervous system responses during perception of masked speech may reflect constructs other than subjective listening effort. *Frontiers in Psychology*, *7*, A263. https://doi.org/10.3389/fpsyg.2016.00263

Gatehouse, S. (1998). Speech tests as measures of outcome. *Scandinavian Audiology*, *27*(4), 54–60. https://doi-org.ezp1.lib.umn.edu/10.1080/010503998420667

Gianakas, S., & Winn, M. (2019). Lexical bias in word recognition by cochlear implant listeners. *Journal of the Acoustical Society of America*, *146*(5), 3372–3383. https://doi.org/10.1121/1.5132938

Gigerenzer, G. (2008). *Rationality for mortals: How people cope with uncertainty*. Oxford University Press.

Griffiths, T., Lieder, F., & Goodman, N. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, *7*, 217–229

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

Hétu, R., Riverin, L., Lalande, N., Getty, L., & St-Cyr, C. (1988). Qualitative analysis of the handicap associated with occupational hearing loss. *British Journal of Audiology*, *22*(4), 251–264. https://doi-org.ezp3.lib.umn.edu/10.3109/03005368809076462

Hsu, N., & Novick, J. (2016). Dynamic engagement of cognitive control modulates recovery from misinterpretation during real-time language processing. *Psychological Science*, *27*, 572–582. https://doi.org/10.1177/0956797615625223

Hughes, S., Hutchings, H., Rapport, F., McMahon, C., & 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 and Hearing*, *39*(5), 922–934. https://doi.org/10.1097/AUD.0000000000000553

Hyönä, J., Tommola, J., & Alaja, A. (1995). Pupil dilation as a measure of processing load in simultaneous interpretation and other language tasks. *Quarterly Journal of Experimental Psycholology*, *48*(3), 598–612. https://doi.org/10.1080/14640749508401407

Ivanova, I., Pickering, M., Branigan, H., McLean, J., & Costa, A., (2012). The comprehension of anomalous sentences: Evidence from structural priming. *Cognition*, *2*, 193–209. https://doi.org/10.1016/j.cognition.2011.10.013

Järvelin, M., Mäki-Torkko, E., Sorri, M., & Rantakallio, P. (1997). Effect of hearing impairment on educational outcomes and employment up to the age of 25 years in northern Finland. *British Journal of Audiology*, *31*(3), 165–175. https://doi.org/10.3109/03005364000000019

Kaandorp, M., Smits, C., Merkus, P., Festen, J., & Goverts, T. (2017). Lexical-access ability and cognitive predictors of speech recognition in noise in adult cochlear implant users. *Trends in Hearing*, *21*, 1–15. https://doi.org/10.1177/2331216517743887

Kadem, M., Herrmann, B., Rodd, J., & Johnsrude, I. (2020). Pupil dilation is sensitive to semantic ambiguity and acoustic degradation. *Preprint available at bioRxiv*. https://doi.org/10.1101/2020.02.19.955609

Kamide, Y. (2008). Anticipatory processes in sentence processing. *Language and Linguistics Compass*, *2*(4), 647–670. https://doi.org/10.1111/j.1749-818X.2008.00072.x

Kerlin, J., Shahin, A., & Miller, L. (2010). Attentional gain control of ongoing cortical speech representations in a "cocktail party." *Journal of Neuroscience*, *30*(2), 620–628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010

Koelewijn, T., de Kluiver, H., Shinn-Cunningham, B., Zekveld, A., & Kramer, S. (2015). The pupil response reveals increased listening effort when it is difficult to focus attention. *Hearing Research*, *323*, 81–90. https://doi.org/10.1016/j.heares.2015.02.004

Koelewijn, T., Versfeld, N., & Kramer, S. (2017). Effects of attention on the speech reception threshold and pupil response of people with impaired and normal hearing. *Hearing Research*, *354*, 56–63. https://doi.org/10.1016/j.heares.2017.08.006

Koelewijn, T., Zekveld, A., Festen, J., & Kramer, S. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing*, *33*, 291–300.

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*(1), 23–34. https://doi.org/10.1111/j.1469-8986.2012.01477.x

Kuchinsky, S., Pandža, N., & Haarman, H. (2016). Linking indices of tonic alertness: Resting-state pupil dilation and cingulo-opercular neural activity. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *LNAI: Vol. 9743. Foundations of augmented cognition: Neuroergonomics and operational neuroscience* (pp. 218–230). Springer. https://doi.org/10.1007/978-3-319-39955-3_21

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163. https://doi.org/10.1038/307161a0

Kuznetsova, A., Brockhoff, P., & Christensen, R. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. http://dx.doi.org/10.18637/jss.v082.i13

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Love, J., Sollman, L., Niehl, A., & Francis, A. (2019). Physiological orienting response, noise sensitivity, and annoyance from irrelevant background sound. *Proceedings of Meetings on Acoustics*, *35*, 1–19. https://doi.org/10.1121/2.0000981

MacGregor, L., Rodd, J., Gilbert, R., Hauk, O., Sohoglu, E., & Davis, M. (2019). The neural time course of semantic ambiguity resolution in speech comprehension. *Journal of Cognitive Neuroscience*, *32*(3), 403–425. http://cognet.mit.edu/journals/journal-of-cognitive-neuroscience/32/3

Mackersie, C., & Cones, H. (2011). Subjective and psychophysiological indexes of listening effort in a competing-talker task. *Journal of the American Academy of Audiology*, *22*(2), 113–122. https://doi.org/10.3766/jaaa.22.2.6

Marrufo-Pérez, M., Eustaquio-Martín, A., & Lopez-Poveda, E. (2019). Speech predictability can hinder communication in difficult listening conditions. *Cognition*, *192*, 103992. https://doi.org/10.1016/j.cognition.2019.06.004

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

McCloy, D., Larson. E., Lau, B., Lee, A.K.C. (2016). Temporal alignment of pupillary response with stimulus events via deconvolution. *Journal of the Acoustical Society of America*. 139, EL57–EL62.

McCloy, D., Lau, B., Larson, E., Pratt, K., & Lee, A. K. C. (2017). Pupillometry shows the effort of auditory attention switching. *Journal of the Acoustical Society of America*, *141*, 2440–2451. https://doi.org/10.1121/1.4979340

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S., & Munro, K. J. (2017). Measuring listening-related effort and fatigue in school-aged children using pupillometry. *Journal of Experimental Child Psychology*, *161*, 95–112. https://doi.org/10.1016/j.jecp.2017.04.006

McGarrigle, R., Munro, K., Dawes, P., Stewart, A., Moore, D., Barry, J., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'. *International Journal of Audiology*, *53*(7), 433–440. https://doi.org/10.3109/14992027.2014.890296

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

McLaughlin, D., & Van Engen, K. J. (2020). Task-evoked pupil response to accurately recognized accented speech. *Journal of the Acoustical Society of America*, *147*, EL151–EL156. https://doi.org/10.1121/10.0000718

Miller, J., Watson, C., Leek, M., Dubno, J., Wark, D., Souza, P., Gordon-Salant, S., & Ahlstrom, J. (2017). Syllable-constituent perception by hearing-aid users: Common factors in quiet and noise. *Journal of the Acoustical Society of America*, *141*(4), 2933–2946. https://doi.org/10.1121/1.4979703

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

Moore, T., & Picou, E. (2018). A potential bias in subjective ratings of mental effort. *Journal of Speech Language and*

Hearing Research, *61*, 2405–2421. https://doi.org/10.1044/2018_JSLHR-H-17-0451

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

Murphy, P., Robertson, I., Balsters, J., & O'Connell, R. (2011). Pupillometry and P3 index the locus coeruleus-norandrenergic arousal function in humans. *Psychophysiology*, *48*, 1532–1543. https://doi.org/10.1111/j.1469-8986.2011.01226.x

Nachtegaal, J., Kuik, D., Anema, J., Goverts, S., Festen, J., & Kramer, S. (2009). Hearing status, need for recovery after work, and psychosocial work characteristics: Results from an internet-based national survey on hearing. *International Journal of Audiology*, *48*(10), 684–691. https://doi.org/10.1080/14992020902962421

O'Neill, E. (2020). *Understanding factors contributing to variability in outcomes of cochlear implant users* [Doctoral dissertation]. University of Minnesota.

Ohlenforst, B., Zekveld, A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N., & Kramer, S. (2017). Impact of stimulus-related factors and hearing impairment on listening effort as indicated by pupil dilation. *Hearing Research*, *351*, 68–79. https://doi.org/10.1016/j.heares.2017.05.012

Peelle, J. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Perry, T., & Kwon, B. (2015). Amplitude fluctuations in a masker influence lexical segmentation in cochlear implant users. *Journal of the Acoustical Society of America*, *137*, 2070–2079. http://dx.doi.org/10.1121/1.4916698

Pichora-Fuller, M. K., Kramer, S., Eckert, M., Edwards, B., Hornsby, B., Humes, L., Lemke, U., Lunner, T., Matthen, M., Mackersie, C., Naylor, G., Phillips, N., Richter, M., Rudner, M., Sommers, M., Tremblay, K., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing*, *37*, 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

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

Popa, A.-B. (2018). Inviting hallucinatory percepts during speech-listening to detect cognitive changes in early psychosis. *Electronic Thesis and Dissertation Repository*, 5854. https://ir.lib.uwo.ca/etd/5854

Reilly, J., Kelly, A., Kim, S. H., Jett, S., & Zuckerman, B. (2019). The human task-evoked pupillary response function is linear: Implications for baseline response scaling in pupillometry. *Behavior Research Methods*, *51*, 865–878.

Reimer, J., McGinley, M., Liu, Y., Rodenkirch, C., Wang, Q., McCormick, D., & Tolias, A. (2016). Pupil fluctuations track rapid changes in adrenergic and cholinergic activity in cortex. *Nature Communications*, *7*, 13289. https://doi.org/10.1038/ncomms13289

Rødvik, A., Tvete, O., Torkildsen, J., Wie, O., Skaug, I., & Silvola, J. (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

Rogalsky, C., & Hickok, G. (2009). Selective attention to semantic and syntactic features modulates sentence processing networks in anterior temporal cortex. *Cerebral Cortex*, *19*, 1786–1796. https://doi.org/10.1093/cercor/bhn126

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

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, O., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience* *7*(31), 1–17. https://doi.org/10.3389/fnsys.2013.00031

Shenhav, A., Botvinick, M., & Cohen, J. (2013). The expected value of control: An integrative theory of anterior cingulate cortex function. *Neuron*, *79*, 217–240

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T., Cohen, J., & Botvinick, 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

Strand, J., Brown, V., Merchant, M., Brown, H., & 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

Strauss, D., & Francis, A. (2017). Toward a taxonomic model of attention in effortful listening. *Cognitive, Affective, and Behavioral Neuroscience*, *17*(4), 809–825. https://doi.org/10.3758/s13415-017-0513-0

Toscano, J., & Allen, J. (2014). Across- and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research*, *57*(6), 2293–2307. https://doi.org/10.1044/2014_JSLHR-H-13-0244

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vaden, K., Kuchinsky, S., Cute, S., Ahlstrom, J., Dubno, J., & Eckert, M. (2013). The cingulo-opercular network provides word-recognition benefit. *Journal of Neuroscience*, *33*(48), 18979–18986. https://doi.org/10.1523/JNEUROSCI.1417-13.2013

Vaden, K., Teubner-Rhodes, S., Ahlstron, J., Dubno, J., & Eckert, M. (2017). Cingulo-opercular activity affects incidental memory encoding for speech in noise. *NeuroImage*, *157*, 381–387. https://doi.org/10.1016/j.neuroimage.2017.06.028

Van der Wel, P. & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. Psychonomic Bulletin & Review, 25, 2005-2015. doi: 10.3758/s13423-018-1432-y

Vogelzang, M., Hendriks, P., & van Rijn, H. (2016). Pupillary responses reflect ambiguity resolution in pronoun processing. *Language, Cognition and Neuroscience*, *31*(7), 876–885. https://doi.org/10.1080/23273798.2016.1155718

Wagener, K., Brand, T., & Kollmeier, B. (1999). Entwicklung und Evaluation eines Satztests für die deutsche Sprache II: Optimierung des Oldenburger Satztests [Development and evaluation of a sentence test for the German language II: Optimization of the Oldenburg sentence test]. *Zeitschrift Audiologie/Audiological Acoustics*, *38*(1), 44–56. https://www.researchgate.net/publication/266735660_Entwicklung_und_Evaluation_eines_Satztests_fur_die_deutsche_Sprache_I_Design_des_Oldenburger_Satztests

Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392–393. https://doi.org/10.1126/science.167.3917.392

Wendt, D., Dau, T., Hjortkjaer, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. Frontiers in Psychology, 7 (345), 1-12. doi: 10.3389/fpsyg.2016.00345

Wendt, D., Koelewijn, T., Książek P., Kramer S., & 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

Winn, M. (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., Edwards, J., & Litovsky, R. (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., & Moore, A. (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., & Teece, K. (2020). Slower speaking rate reduces listening effort among listeners with cochlear implants. *Ear and Hearing*. *Advance online publication*. https://doi.org/10.1097/AUD.0000000000000958

Winn, M., Wendt, D., Koelewijn, T., & Kuchinsky, S. (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*, 1–32. https://doi.org/10.1177/2331216518800869

Zekveld, A., Koelewijn, T., & Kramer, S. (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., Kramer, S., & Festen, J. (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