Cognitive and physiological measures of listening effort during degraded speech perception: Relating dual-task and pupillometry paradigms

Sarah Colby[1] & Bob McMurray[1]

[1]Department of Psychological & Brain Sciences, University of Iowa,

Iowa City, IA

Corresponding author:
Sarah Colby
G60, Psychological & Brain Sciences Building
340 Iowa Ave.
Iowa City, IA
52242
E-mail: sarah-colby@uiowa.edu

Conflict of Interest: The authors have no conflicts of interest to report.
Funding Statement: This work was supported by NIH grant P50 000242 awarded to Bruce Gantz and Bob McMurray, and DC 0008089 awarded to Bob McMurray.

Abstract

Purpose: Listening effort is quickly becoming an important metric for assessing speech perception in less-than-ideal situations. However, the relationship between the construct of listening effort and the measures used to assess it remain unclear. We compared two measures of listening effort: a cognitive dual task and a physiological pupillometry task. We sought to investigate the relationship between these measures of effort and whether engaging effort impacts speech accuracy.

Method: In Experiment 1, 30 participants completed a dual task and pupillometry task that were carefully matched in stimuli and design. The dual task consisted of a spoken word recognition task and a visual match-to-sample task. In the pupillometry task, pupil size was monitored while participants completed a spoken word recognition task. Both tasks presented words at three levels of listening difficulty (unmodified, 8-channel vocoding, and 4-channel vocoding) and provided response feedback on every trial. We refined the pupillometry task in Experiment 2 (n=31); crucially, participants no longer received response feedback. Finally, we ran a new group of subjects on both tasks in Experiment 3 (n=30).

Results: In Experiment 1, accuracy in the visual task decreased with increased signal degradation in the dual task, but pupil size was sensitive to accuracy and not vocoding condition. After removing feedback in Experiment 2, changes in pupil size were predicted by listening condition, suggesting the task was now sensitive to engaged effort. Both tasks were sensitive to listening difficulty in Experiment 3, but there was no relationship between the tasks and neither task predicted speech accuracy.

Conclusions: Consistent with previous work, we found little evidence for a relationship between different measures of listening effort. We also found no evidence that effort predicts speech

accuracy, suggesting that engaging more effort does not lead to improved speech recognition. Cognitive and physiological measures of listening effort are likely sensitive to different aspects of the construct of listening effort.

Cognitive and physiological measures of listening effort during degraded speech perception: Relating dual-task and pupillometry paradigms

Research on spoken language comprehension in adverse listening conditions has shown increasing interest in the construct of *listening effort*, the recruitment of additional cognitive resources for speech recognition (Ohlenforst et al., 2017; Wagner et al., 2016; Wendt et al., 2016). Listening effort has been characterized in several different ways, but there is broad agreement that it requires the engagement of cognitive resources towards completing a listening goal (Francis & Love, 2019; Pichora-Fuller et al., 2016). Attention and working memory are most often invoked as the cognitive resources that are engaged in effortful listening tasks, and these presumably support speech perception. It may be particularly critical to understand listening effort for two broad goals.

First, at a clinical level, effort may help understand differences among individuals in their subjective outcomes after hearing intervention. Two adults might perform similarly in in-clinic or in-laboratory speech perception measures, but one may report subjectively that speech perception in the real world is more difficult. Measures of effort in these laboratory tasks may help us understand this disconnect. Alternatively, two listeners may differ in performance when trying to perceive speech in a noisy or reverberant situation, not because they have different levels of skill at dealing with degraded input, but because they devote more or less effort.

Second, effort raises a scientific puzzle for theories of speech perception. Listening effort is typically characterized in terms of cognitive resources. However, no models of speech perception or word recognition include any notion of "resources" (TISK, Hannagan et al., 2013; TRACE, McClelland & Elman, 1986; C-CuRE, McMurray & Jongman, 2012; Shortlist B, Norris & McQueen, 2008). For normal hearing adults in quiet—which these theoretical models were developed to account for— this is a reasonable assumption. However, in more challenging

situations, people clearly engage extra cognitive effort (Koelewijn et al., 2015; Wu et al., 2016; Zekveld & Kramer, 2014). This presumably has the goal of compensating for less-than-ideal listening environments; however, it is unclear whether this actually improves speech perception, or whether the additional resources are devoted to processes that are indirectly related to accuracy, like error monitoring. There are no theories that explain how these resources mechanistically improve perception. Thus, problems like speech-in-noise point to a major limitation of current models, and an opportunity to link these theoretical models of language processing into domain general cognitive constructs.

These goals may interact in complex ways. It is widely agreed that under ideal listening conditions speech is usually processed automatically. However, engaging additional effort can lead to fatigue (which would not be experienced with more automatic processing). As a result, the degree to which effort is required for a listener may have downstream impacts on one's motivation and mental fatigue in engaging with speech in the real world, a critical problem for hearing-impaired listeners (Hornsby, 2013; Zekveld et al., 2010). As a result of this cascade, listening effort might be an important predictor of success after hearing remediation. This is not captured in standard clinical tests, but a better understanding of effort may lead to assessments that can target this source of variance. Moreover, understanding and alleviating the need for increased listening effort is important for improving the quality of life of listeners who routinely encounter challenging listening. However, without understanding whether and how effort improves perception, it is not be clear how and when to intervene.

The Framework for Understanding Effortful Listening model (FUEL; Pichora-Fuller et al., 2016) offers some ways to put these pieces together. Under FUEL, the amount of effort engaged for a task is dictated by an individual's cognitive reserves, the difficulty of the task, and

their motivation to complete the task. This model suggests that one's motivation for a task are unique to the individual, with differing impacts on real-world outcomes or speech in noise perception. Peelle (2018) provides an extensive overview of the cognitive resources that are called upon in challenging listening, including working memory and attention, while acknowledging how these resources are likely called upon in different ways depending on situational demands. Several recent models propose a framework under which to assess listening effort (Herrmann & Johnsrude, 2020; Strauss & Francis, 2017). While this work offers a clear picture of the relevant components of the system and how they may vary, it remains unclear what these resources are, what cognitive systems they derive from, and how they work to improve speech perception.

Evidence from neuroimaging points to neural mechanisms that underly effort. This can identify systems that work together to solve challenging listening situations. For instance, pupil dilation—a common measure of effort—has been correlated to activation in bilateral superior temporal gyrus (STG) during challenging speech perception tasks (Zekveld et al., 2014), suggesting that physiological markers of effort can reflect activation in brain regions tied to speech processing. Pupil size itself is regulated both sympathetically and parasympathetically, with cognitive demands resulting in pupil dilation through the parasympathetic system (see Francis & Love, 2019). This work highlights the complex interaction between systems supporting effort and those responsible for speech recognition.

However, the first step in understanding the interaction between effort and speech recognition is the ability to measure effort itself. A variety of methods are used, including self-report, psychological, and physiological measures. It is currently not clear how these methods relate, and whether they capture the same phenomena because few studies have used multiple

methods. While the psychological and physiological measures are often treated as specific measures of effort, they are likely also sensitive to other factors like motivation, anticipation of reinforcement, self-monitoring of performance, or other non-auditory cognitive functions (Ahern & Beatty, 1979; Koelewijn et al., 2015, 2018). These factors may not impact speech perception directly (in the same way that engaging working memory or cognitive control is thought to) but are not entirely independent of it. The field is taking note of this and issues regarding the construct validity of listening effort measures are beginning to be addressed (Strand et al., 2021).

The present study builds on this work. We examined two common methods of measuring listening effort—the dual-task paradigm and pupillometry—in similar tasks with the same stimuli and in the same subjects. The most important question we address here is how these measures relate to each other. However, as exploratory goals we also ask how these relate to both speech perception accuracy, and (across experiments) to the anticipation of reinforcement.

Measures of Listening Effort

Dual Task Performance. In dual tasks, the listener simultaneously process speech and perform an unrelated—but cognitively demanding—task. The assumption is that effort is a domain-general resource which can be deployed to auditory, visual or cognitive tasks. If effort is used for speech perception, then it must be drawn from the same pool of finite cognitive resources used for the unrelated task. As a result, there should be a performance decrement (either to accuracy or response time) in the secondary non-speech task when the listener is simultaneously doing the speech task (or when the speech task gets more difficult). Critically, by focusing on performance in the non-speech task, these tasks can partially dissociate effort from speech perception performance.

A variety of tasks has been used for both the primary speech task and secondary task. The speech task can be anything from sentence recognition (Ward et al., 2017), to serial recall of words (Hornsby, 2013), or speech discrimination (Mitterer & Mattys, 2017). Secondary tasks are typically an unrelated non-speech task, such as visual search (Mattys et al., 2014; Mattys & Wiget, 2011; Mitterer & Mattys, 2017) or visual monitoring (Hornsby, 2013; Mitterer & Mattys, 2017; Ward et al., 2017). A common paradigm is to manipulate the difficulty of the speech task, while holding the non-speech task constant at some level. Engagement of effort is then monitored by observing how performance in the non-speech task changes as a function of the difficulty of the speech task (see Gagné et al., 2017 for review). For example, Ward et al. (2017) assessed sentence recognition at several levels of degradation with a simultaneous visual monitoring task in participants had to detect repeated images in a sequence. Older adults showed greater declines in performance on the visual task while simultaneously recognizing degraded speech, suggesting they engage additional effort compared to younger adults.

There are drawbacks to dual task designs. First, they must use a second task to either impose a cognitive load on the primary speech task or, as in our design, to track changes in performance as the demands of the speech task increase. This means that effort must be inferred across tasks or conditions. That is, effort is measured in relation to two tasks, rather than just the effortful task itself. This is not an issue if the question primarily concerns a difference among conditions. However, when using dual task measures as an individual-level measure (e.g., to assess differences in cognitive resources that individuals engage), this introduces more variability to the measures, since the individual measure must rely on difference scores (which have twice as much variability). This is a concern for already variable populations like listeners with hearing impairment. Second, dual tasks require many trials to adequately calculate

difference scores and are thus can be rather time consuming. Nonetheless, the dual-task paradigm has a strong grounding in a cognitive literature on dual task performance, and thus has clear construct validity.

Pupil Dilation. Task-evoked changes in pupil size have also been used as a measure of task engagement. Here, increased engagement is thought to require greater devotion of cognitive resources. Generally, pupil size is regulated by the interaction of the sympathetic and parasympathetic nervous systems. Cognitive demand results in the inhibition of the parasympathetic nervous system, which dilates the pupil (Steinhauer et al., 2004; Zekveld et al., 2010). While completing a demanding task, pupil size will increase until a decision has been made and then it decreases (van der Wel & van Steenbergen, 2018).

Pupillometry studies have mostly used sentence recognition tasks. These are likely preferred (over single word tasks) in part because the longer stimulus offers a more time over which to observe the dynamics of pupil size changes. These studies employ various ways to increase difficulty, including presenting speech in noise (Koelewijn et al., 2017; Lau et al., 2019; Ohlenforst et al., 2017; Zekveld & Kramer, 2014) or using noise-vocoded speech (Winn, 2016; Winn et al., 2015; Winn & Moore, 2018). For example, Winn (2016) used pupillometry to ask if sentence context reduces listening effort in cochlear implant (CI) users and normal hearing (NH) listeners (hearing vocoded speech). Both groups of subjects exhibited a larger change in pupil size to low predictability sentences than to high predictability sentences, suggesting that predicable contexts reduced the effort required.

Pupillometry comes with several challenges. One drawback that changes in pupil size may reflect multiple underlying cognitive processes. While the physiological mechanisms that dilate and constrict the pupil are established, these mechanisms are sensitive to many different

psychological factors. For example, during studies of listening effort, changes in pupil size can also be influenced by attention (Baldock et al., 2019; Koelewijn et al., 2014, 2015, 2017), fatigue (McGarrigle et al., 2017), and reward (Koelewijn et al., 2018; Zekveld et al., 2019). These factors may be independent of the cognitive resources devoted to the task (perhaps reflecting something like arousal), or they could reflect aspects of effort that do not contribute to speech intelligibility. For example, if pupil changes are driven by anticipation of reward, this may reflect a reversal of the typically assumed causal pathway. The standard assumption that increased effort leads to better performance as more resources are devoted to speech perception. However, what if pupil dilation reflects subjects' monitoring of their own performance, with greater perceived challenge leading to greater dilation? Here, changes performance are what drive changes in pupil dilation!

Clinically, this distinction matters. Under the standard view, cognitive training to alter the allocation of effort may be a promising avenue to improve outcomes; however, under the latter view, we should focus increase signal quality (the periphery) to reduce effort. Disentangling these causal pathways is also crucial for building effortful processing into mechanistic models of speech perception: under the standard assumption, effortful processing must be engaged directly in the processing pathway; while under a reversed model, speech perception itself can largely be autonomous, and effort need only be aware of the outcome. This highlights the importance of carefully designed and controlled experiments to allow for conclusions to be drawn from pupillometry results.

Toward a more unified view of listening effort.

Most studies using either approach have focused on documenting which listening

conditions lead to greater effort (e.g., greater SNR, hearing loss, etc.). However, to move toward a more unified view, we must address two questions. First, and most importantly, we ask if different measures of effort are correlated with each other. Second, it is still unclear whether effort plays a mechanistic role in speech perception: does exerting effort make speech perception more accurate? The latter question is much less straightforward to answer as it is not clear how to experimentally manipulate effort to observe the results on performance. Thus, the present study was primarily designed to address the former by matching task designs across two measures of effort. We attempt to address the second question in a more exploratory way by also examining the relationship between effort and speech recognition accuracy.

Cross-measure comparison. Several studies have tried ask if different measures of listening effort relate to each other and to the overall construct by assessing the relationship between self-reported feelings of effort or fatigue and dual-task or pupil measures. These have shown inconsistent results. Subjective reports often trend in the same direction as other measures (Koelewijn et al., 2014; Picou et al., 2011; Picou & Ricketts, 2018). For example, listening conditions that are reported as feeling more effortful will also elicit a larger pupil response (Koelewijn et al., 2014). However, other studies find little relationship between subjective and objective measures of listening effort (Feuerstein, 1992; Lau et al., 2019). For example, Lau et al. (2019) found no relationship between pupillometry and a subjective measure of effort during word and sentence recognition in noise.

There is recent evidence that dual tasks and pupillometry can provide converging measures of the effort expended in a speech perception task, with both pupil response and dual task behavior reflecting increased effort to accented speech (Brown et al., 2020). Similarly, Karatekin et al. (2004) measured pupil response during a dual task, and found increased pupil

size was found in dual-task trials compared to single-task trials, even when accuracy was not affected (but response time was). Thus, pupil size and dual task measures generally respond to similar experimental manipulations.

However, other work has directly related multiple psychological and physiological measures and found little correlation (Alhanbali et al., 2019; Strand et al., 2018). The absence of a relationship between measures suggests either that different methods are likely tapping different aspects of listening effort, or that listening effort itself is not driven by a single cohesive construct. There are still open questions about the relationship between measures of listening effort, particularly between pupillometry (what is rapidly becoming the go-to measure), and dual-task measures (with stronger grounding in cognitive theory).

For instance, Alhanbali et al. (2019) collected seven measures of listening effort and conducted a factor analysis. Different measures of effort generally loaded onto different components, suggesting listening effort is not a cohesive construct. However, their methods may have constrained this conclusion. They used a digit recall task in which participants heard six spoken digits in noise, and after a brief delay, had to recognize whether a new digit was present in the list. While completing this task, pupillometry, EEG, and skin conductance were recorded. However, these stimuli represent a limited, closed set of highly recognizable items. Such a limited set of items calls into question the validity of generalizing these findings to how effort influences speech perception more broadly.

Strand et al. (2018) collected several measures of listening effort (including cognitive, physiological, and subjective) on a large sample (n=111). While some cognitive measures were correlated with each other, these were generally not related to the physiological and subjective measures. However, they did not match the stimuli for pupillometry and dual tasks. Both dual

tasks required listeners to hear and repeat words (the speech perception task), while concurrently responding to visually-presented numbers (Sarampalis et al., 2009) or judging whether the word was a noun (Picou & Ricketts, 2014). Meanwhile, their pupillometry task used a listen-and-repeat speech perception task with sentences (Zekveld & Kramer, 2014). However, the magnitude of change in pupil size is not the same for sentence and word recognition (Lau et al., 2019). As a result, the attempt to compare measures of listening effort across these speech perception tasks might introduce additional variance that obscures a potential relationship.

Additionally, the sentence recognition task used in both studies (and many others) engage more domain general resources than single word tasks. Sentence recognition and production (as in Strand et al., 2018) clearly require working memory (Just & Carpenter, 1992) and executive function (Novick et al., 2005). Moreover, the digit recall task (Alhanbali et al., 2019) requires working memory. Working memory and executive function are domain general and resource limited (Ma et al., 2014). In both cases, enhanced effort could be applied not to the processes of speech perception, but rather to these cognitive processes. Thus, it is unclear whether the increased effort really reflects effortful listening, or something downstream. This is not a problem with respect to the issue of correlating measures of effort more broadly; however, it is a big issue for understanding the mechanistic role of effort in speech perception.

Thus, there remains a need to relate listening effort measures that are better matched to each other and which better isolate word recognition from other cognitive processes. The present study addresses this first issue by designing a dual task and pupillometry task that use the same stimuli and the same basic task. We used a closed-set single-word recognition task that does not depend on speech production for a response. This allows us to focus on the effort required for recognizing speech, without interference from domain general cognitive processes involved in

sentence processing, working memory, or speech production.

The link between listening effort and speech perception accuracy. Most listening effort experiments show changes in effort with increasing auditory demand. This does not address the underlying question of whether allocation of effort contributes to success at speech recognition in challenging situations. A secondary (and more exploratory) aim of the current study seeks to understand how measures of effort relate to speech perception accuracy as part of a longer term-goal of using both measures as a method for investigating individual-level differences in hearing impaired populations (e.g., to predict other outcomes).

One consistent finding across both pupillometry and dual tasks is that listening effort typically shows a 'U-shaped' curve with respect to difficulty. That is, effort (as indicated by peak pupil size or dual-task interference) increases as the task becomes more difficult, until the point where the task is too difficult and it decreases (c.f., Wu et al., 2016; Zekveld & Kramer, 2014). The fact that this is seen with both pupil and dual-task measures lends credence to the idea that both dual-task and pupil measures at some level capture the same construct.

However, the drop-off in performance at high levels of difficulty raises important questions. It suggests that if a task is too cognitively demanding, subjects might show what appears to be the same level of effort as during a task that is not demanding. Indeed, the participants who admitted to giving up during low intelligibility sentences also had smaller pupil sizes during those trials (Zekveld & Kramer, 2014). This suggests that these measures do not solely reflect the demand imposed by task (if they did, they would increase monotonically).

Several studies manipulate intelligibility to investigate the amount of effort engaged to accomplish a given level of accuracy (as in establishing evidence for the 'U-shaped' curve described above, Wu et al., 2016; Zekveld & Kramer, 2014). Intelligibility does not predict

maximum pupil size independent of listening condition (Winn et al., 2015). This seems to imply that after accounting for overall difficulty, subjects who exert more effort do not show better accuracy. There is, in fact, a growing body of evidence that speech accuracy and listening effort are not inherently tied together and that increases in effort do not necessitate improvements in performance (Mackersie & Cones, 2011; Sarampalis et al., 2009; Winn & Teece, 2020).

As a secondary goal, the present study sought to examine the relationship between speech recognition accuracy and effort. As much of the past work was done with sentences, some of the variation in effort may have been due to higher level cognitive processes, making it more difficult to see a subtle perceptually-driven effect. The current study adopts a similar logic to Winn et al. (2015) in correlating effort measures with accuracy over and above listening condition to examine the relationship (if any) between speech accuracy and engaged effort. This is necessarily an exploratory (and correlational) approach.

Moreover, the *slope* of the pupil response over time –from baseline to peak pupil dilation—might be more theoretically tied to speech recognition accuracy. The onset slope from baseline to peak pupil dilation over time is thought to be tied to auditory processing demands (Winn et al., 2015), while the recovery slope after the peak (steeper return to baseline after peak) has been linked to aptitude (Ahern & Beatty, 1979; Bianchi et al., 2016). Bianchi et al. (2016) found that musicians have a quicker recovery than non-musicians while discriminating pitch than non-musicians, suggesting skill at a task can help reduce effort faster when it is no longer needed. However, the slope (onset or offset) has not yet been correlated on a between-subject basis with speech perception accuracy. Thus, the present study includes the timing of the pupil size peak (latency to reach maximum pupil size) in our analyses to investigate how the timecourse of changes to pupil size reflect aspects of effort and relate to accuracy.

The Present Study

The present study directly relates results from closely matched dual-task and pupillometry paradigms to examine the relationship between these measures of listening effort and between effort and accuracy. This current set of experiments were intended as preparatory work for a larger individual differences study on cochlear implant users; in that study, the primary measure would be a Visual World Paradigm (VWP) measure of lexical activation dynamics. Thus, many design choices were motivated by a need for tasks that are aligned to each other and to the VWP, and by our primary interest in between-subject variability.

We used noise vocoding as our difficulty manipulation because of our ultimate goal of testing cochlear implant users. Difficulty was manipulated by using 4- and 8-channel vocoding along with unmanipulated speech. This was based on prior work in the VWP which shows a moderate (~80 msec) delay in lexical access for 8-channel vocoding (Farris-Trimble et al., 2014), and a large (~200 msec) delay for 4-channel (McMurray et al., 2017). We expected accuracy to be somewhat high, even in 4-channel vocoding (McMurray et al., 2017 report 80% accuracy using similar stimuli). However, we note that 80% accuracy in a closed set task such as this one translates to much lower accuracy in an open set task (see Figure 1 of Clopper et al., 2006) which are more common in this literature.

We tested listeners in a closed-set task on single words. The use of words in isolation does away with the domain general resources required to process sentences to focus on the effort engaged for speech-specific processing, and the closed-set task eliminates any speech production demands. This task was modeled after the VWP (Farris-Trimble & McMurray, 2013) in which a single word is heard, and the listener uses a mouse to select a picture from an array that contains the target and several similar sounding competitors. This task much more closely matches the

some of the demands of real-world word recognition as it requires listeners to map incoming speech to semantic referents (rather than to orthographic word forms, or to articulatory plans). However, note that unlike the VWP, in both tasks, subjects were not aware of the response options in advance (e.g., there was no pre-scan)—consequently, at the time at which they were engaging effort these were functionally open-set (not closed-set tasks).

We developed two variants that were customized for either a dual-task format or for pupillometry. Unlike many dual-task implementations, ours relies on changes in accuracy—not reaction time—under conditions of high load. This is primarily because we expected the mouse response to add significant variability to the measure; and switching to a button press would add significant cognitive load as participants needed to track which item corresponded to which button on each trial.

Using these tasks, we asked if the performance decrement with vocoded speech in the dual-task paradigm correlates with maximum pupil size and maximum pupil time. Secondarily, we related both measures to speech perception accuracy in the same task. In Experiment 1, subjects completed roughly standard versions of both tasks using the same stimuli, and approximately similar task demands. A close analysis suggested that the original pupillometry task may have been incidentally influenced by the anticipation of response feedback. Thus, in Experiment 2, we redesigned and validated the pupillometry task, and in Experiment 3 ran a new group of subjects on both measures.

Methods

For this and all experiments, we have reported all of the measures and conditions that were run, and any data exclusions that were employed.

Participants

Forty-two participants were recruited from the University of Iowa community. Subjects had normal or corrected-to-normal vision and had no reported hearing loss. They received either partial course credit or monetary reimbursement for their participation. Five participants did not return to complete the second session and five did not finish the tasks during their visits. One participant was excluded because of age, and one because of low accuracy. This left full data sets from 30 participants. We set 30 subjects as our target sample size. A minimum detectable effect with N = 30, alpha = 0.05, and a power of 0.8 is $r \geq 0.42$. Given that we were trying to correlate our two effort measures, a smaller effect would not have been of interest. Data collection continued until we believed we had data from 30 subjects in both tasks.

Procedure

Participants completed two sessions one week apart. The dual task was completed during the first session, and the pupillometry task was completed during the second session. We used a fixed—rather than counterbalanced—task order for two reasons. First, in the dual-task, accuracy was the dependent variable, whereas the pupillometry task used pupil dilation. Consequently, we didn't want participants to have any familiarity with vocoding (which would affect accuracy) before completing this task. Second, this was a correlational study which relies on systematic variance between subjects. If we had counterbalanced the task order, some of the between subject variance would be due to differences in order (not differences in skill or effort in processing vocoded speech); this would depress the effect size.

Dual task.

There are two common variants of the dual task. First, speech perception task can be held constant, while the demands of the secondary (visual) task are manipulated (e.g., Mattys & Wiget, 2011); such designs are appropriate when the question is whether load impacts speech perception performance (since the primary dependent variable is speech perception). Alternatively, we can manipulate the demands of the primary speech task, and examine changes in a constant secondary task (Francis & Love, 2019, section 3.1.3). This is more appropriate when the goal is to estimate the amount of load that speech perception requires (since the dependent variable is not directly tied to speech perception performance). We adopted the latter, assuming that as the speech task becomes more demanding, performance on the visual task will suffer (increased errors, slower response time) because of the finite resource pool available for both tasks.

Thus, the dual task was built from a baseline – visual only – task in which subjects matched complex visual forms. This baseline task was then turned into a dual task in which subjects simultaneously did a speech perception task. Our primary measure was thus performance on the visual task, which was examined as a function of the difficulty of the speech task. The unmodified speech should have been processed nearly automatically, and therefore would use few resources, where the higher levels of vocoding should introduce more decrement to the visual task performance.

During the dual task, participants were told that they would encounter several trial types: (1) grid matching only, (2) preview a grid and hear a word, respond to the grid and (3) preview a grid and hear a word, respond to the word. During trials where participants had to attend to both a grid and a word, they did not know which they would respond to until the response screen appeared. Thus, they had to prepare both responses. On all trials, participants received both

visual and auditory feedback on their response, indicating whether they made a correct or incorrect choice. A green box would appear around the selection and a ding sound would play for a correct response and a red box would appear with a buzzer sound for an incorrect response. This feedback was intended to keep participants engaged in this rather long and repetitive task.

Five practice trials exposed participants to each of the trial types and to vocoded words. An experimenter would answer any procedural questions after the practice trials before the main task proceeded. The dual task took approximately an hour to complete.

Baseline Task. The baseline task was a visual match-to-sample task (Figure 1A). Participants matched a grid pattern to one of four options at a 1500 msec delay. The target grid was visible for 1000 msec, then there was a 1500 msec delay, and the four grids from a set (one target, one competitor, two distractors) appeared on the screen. The participant then selected the grid that matched the previewed target. The motivation for this particular task design was twofold: we wanted to equate task demands across the speech and visual trials and this would permit easier interleaving of trials. Presenting a visual target to match at a delay imposed the same structure of a speech recognition trial (hear a word, choose the target) on the non-linguistic task. There were 75 baseline trials randomly interspersed among the dual task trials.

Dual Task. During dual task trials, the target grid appeared on the screen for 1000 msec. 300 msec into the grid preview, an auditory word played. After a 1500 msec delay (from the offset of the grid), the response choices (either four grids or four pictures from the target word's set) appeared. Participants did not know whether they would respond to the grid or to the target word until they saw the response options. Words were presented unmodified or at one of two levels of vocoding.

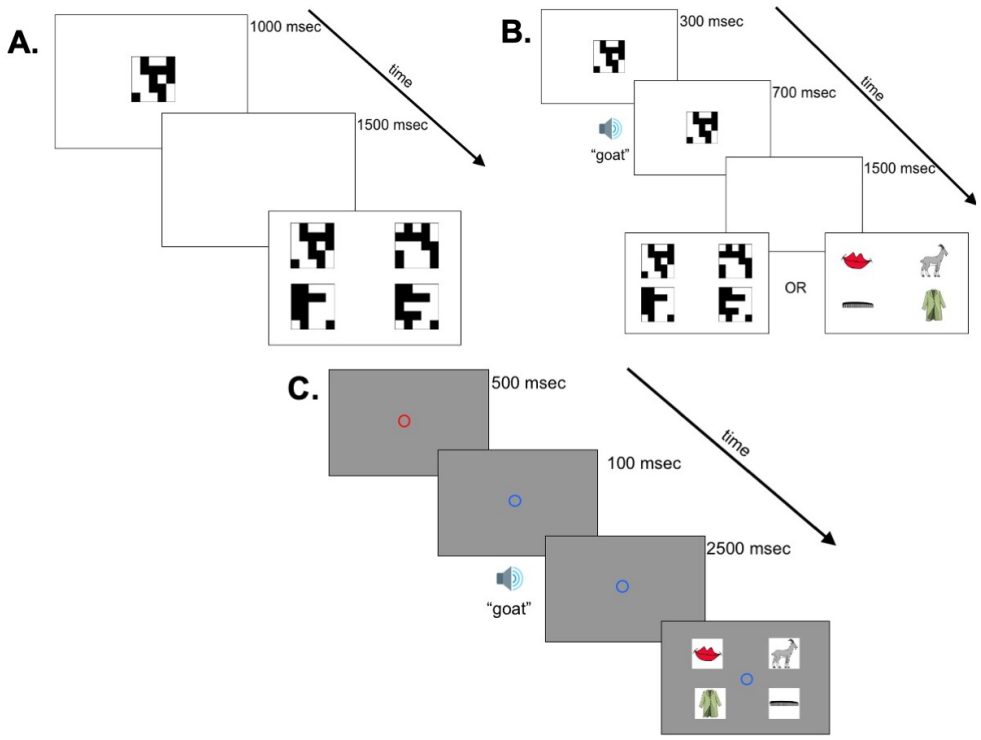

Figure 1. (A) Progression of the baseline trials, (B) the dual-task trials for the dual-task paradigm, and (C) the pupillometry trials.

Design. Each word was heard in every auditory condition (3 levels of degradation) and each response condition (auditory or visual). Additionally, one word from each item set was chosen at random for an additional trial in each condition, such that participants would not be able to eliminate response options just because they had heard that target already.

This resulted in 450 dual task trials (15 item sets × 5 items × 3 listening conditions × 2 response conditions), with an additional 75 baseline trials. Visual (grid) response trials and auditory (speech) response trials were mixed with baseline trials, but level of speech degradation (unmodified, 8-channel, 4-channel) was blocked. This was done to ensure that listeners knew how difficult to expect the trial before it began so that they could deploy effort accordingly. Blocks were randomized so that each listening condition appeared once in the first half of the experiment and once in the second half (half of the blocks tested 87 trials, the other half tested 88

trials).

Pupillometry.

The pupillometry task was a word recognition task, similar to the VWP and to the dual task trials (without the visual task; see Figure 1C). At the beginning of a trial, a red circle appeared at screen center. After a 500 msec delay, the circle turned blue, which signaled that the participant could click on the circle to play the word. The blue circle remained on the screen for 2500 msec, and participants were instructed to remain looking at the fixation circle while it remained on the screen (i.e., during the 2500 msec before the onset of the response screen). Nothing else was presented on the screen during this period. This was intended to provide long measurement period in which the screen was largely empty, and eye-movements were minimized, but that subjects were actively processing the word. After 2500 msec, four pictures corresponding to the item set for that trial appeared, and participants clicked on the image that matched the word. The screen background for the entire task was set to grey (RBG 150, 150, 150) so that the pupil would not be overly constricted by a bright white screen.

Before the experiment, there were four practice trials after which the experimenter answered any questions. The pupillometry task took about 45 minutes.

Design. Each item served as the target at least once and two items from each set were randomly chosen for additional repetitions as the target. This led to a total of 360 trials (4 words/set × 20 sets × 1.5 repetitions[1] × 3 listening conditions). Trials were blocked by level of degradation (unmodified, 8-channel, 4-channel), such that a block of each condition would appear in the first and second half of the experiment (60 trials/block). Participants again received

[1] To get 1.5 repetitions of the stimuli, each item from a set was the target once and then 2 items from a set were randomly chosen to be the target a second time.

visual feedback on their responses to remain consistent across tasks.

Pupillometry Measures and Calibration. Prior to the pupillometry task, an experimenter calibrated an Eyelink 1000 desktop-mounted eyetracker using a 9-point calibration. The eyetracker was set to sample at 250 Hz. To measure an individual's dynamic pupil range, participants next saw a white screen for 15 seconds, then after a 500 msec delay, a black screen for 15 seconds. During this dynamic range estimation, participants were instructed to keep their gaze relatively still and relatively centered on the screen. To help with this, a 500 × 500 pixel box outlining the center of the screen was visible in light grey on the white screen or dark grey on the black screen. These measures were later used to normalize pupil size to each individual's range.

Data Processing. Pupil size was processed using a newly created version of EyelinkAnal (version 4.11; McMurray, 2019). Pupil size was first normalized to an individual's dynamic range (Ayasse et al., 2017; Winn et al., 2018). Maximum pupil size (*max* in Eq 1) was taken as the average of the last 25% of the black calibration screen and minimum pupil size (*min*) was the last 25% of the white calibration screen. Pupil size was scaled by the formula in (1).

$$\text{Scaled pupil size} = (\text{raw pupil size} - \text{min})/(\text{max} - \text{min}) \times 100 \quad (1)$$

Subjects were centrally fixating during 66.7% of the samples during the unmodified speech condition, 66.3% of the 8-channel vocoding condition, and 64.9% of the 4-channel vocoding condition. We did not attempt to interpolate over blinks as we found that this created artifacts in the data. Instead, the frames of a blink were simply treated as missing data in computing the average at that time. Blinks were extended by 50 msec on both sides to get rid of the steep drop off and rise in pupil size that results from the eye closing and opening. This resulted in 16.8% of samples being dropped from the unmodified speech condition, 18.9% from the 8-channel vocoding condition, and 21.2% from the 4-channel vocoding condition.

Lastly, each trial was baselined to an average of the 300 msec preceding the onset of the auditory stimulus. Change in pupil size (from baseline) was then averaged for each condition for the 2500 msec between the onset of the auditory stimulus and the appearance of the response screen. This captures the time during which the participant is listening to the target word but before any eye movements would be launched to an image. We did not analyze pupil size once the screen changed to avoid dealing with the luminance changes and eye movement artifacts. Maximum pupil size was extracted from the average timecourse for each condition and participant for the first 2000 msec after the onset of the target. Maximum pupil time was also extracted from each condition as the average of the timestamps at which pupil size was at least 98% of its maximum size.

Stimuli

Speech Stimuli. Speech stimuli were the same across both tasks. Twenty item sets from Farris-Trimble, McMurray, Cigrand, and Tomblin (2014) were used in the pupillometry task. A subset of 15 item sets was used in the dual task. Each item set was comprised of four words: a target, cohort competitor, rhyme competitor, and unrelated distractor (e.g., *sandal*, *sandwich*, *candle*, *necklace*). The items were noise vocoded at two different levels of difficulty (8-channel and the more difficult 4-channel) using AngelSim (version 1.07.01; Emily Shannon Fu Foundation, 2012).

Visual Stimuli. Visual depictions of the words were the same across both tasks. We used 80 clipart images corresponding to each of the words, as in Farris-Trimble et al. (2014).

For the dual task, 5 × 5 black and white square grid patterns were created in Matlab (MATLAB 2017a, 2017). 525 unique target grids were created randomly, and three additional

grids were permuted from each target. One competitor differed from the target grid in one white and one black square, and the remaining two distractors were randomly rearranged from the target while maintaining the overall amount of black and white within the grid. All images were sized to 300 × 300 pixels. Stimuli are available at https://osf.io/qs6b9/.

Results

We started by examining each task individually as a function of listening condition to document that degradation had the predicted effect on effort. We then followed up on this with a communality analysis to ask how listening condition and accuracy independently predict effort. Next, we turned to our first major question to determine if the two tasks are related. Finally, we related both effort and condition to accuracy to ask if effort improves speech perception. Data and R scripts to recreate all analyses are available at https://osf.io/qs6b9/.

Effect of Condition on Effort.

Dual Task. Mean accuracy for each trial type is reported in Table 1 and Figure 2. As expected, on the speech response trials listeners showed a large effect of degradation, performing above 99.7% correct with unmodified speech, but falling to 75% in the 4-channel vocoding. A similar effect was seen in the visual task. At baseline and with unmodified speech in the dual task participants were at about 83%, but this fell to 79% with 4-channel vocoding.

To analyze this statistically, responses from were coded as correct (1) or incorrect (0) and used as the dependent variable in a logistic mixed effects models with listening condition as a predictor. Separate models were run for the grid-response—our primary measure of effort—and the speech-response trials—to evaluate the effect of the difficulty manipulation (vocoding

Table 1. *Mean accuracy by trial type and listening condition in the dual task.*

		Baseline	Unmanipulated Speech	8-channel	4-channel
Single task	Grid response	83.4 %			
Dual task	Grid response		83.1 %	80.1 %	79.5 %
	Speech response		99.7 %	93.2 %	75.4 %

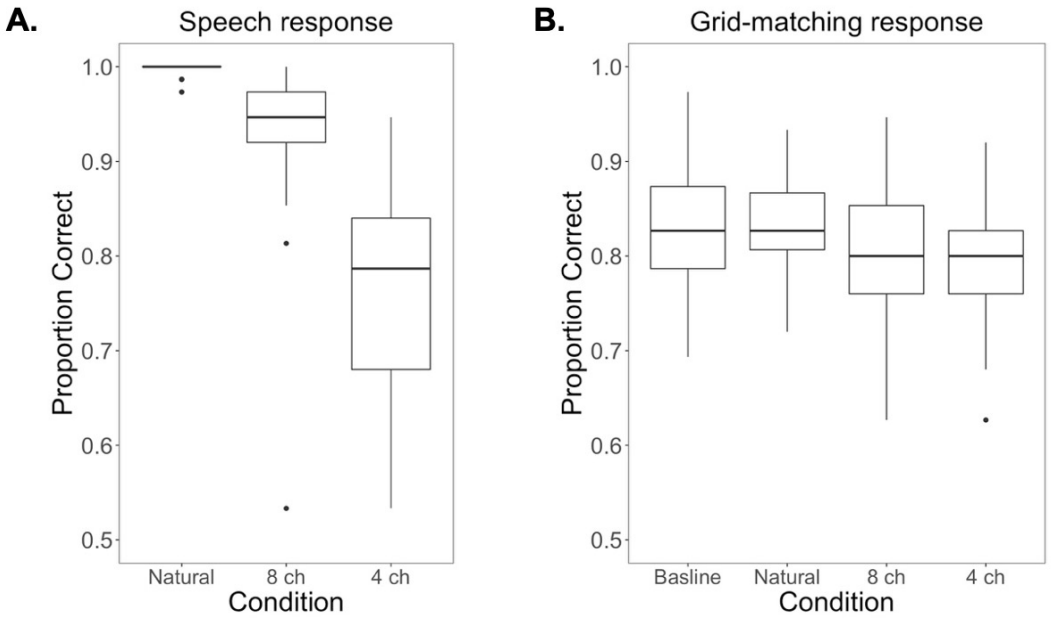

Figure 2. Proportion of correct trials in the dual task by each trial type. (A) Dual-task trials where participants were responding to auditory target words, while (B) is all trials where participants were responding to the visual grids. There were only baseline trials (i.e., single-task trials) for grid response trials.

condition) on intelligibility. We only analyzed responses from the dual task trials, not the single-task baseline trials, to isolate the impact of listening condition and not any change in behaviour that might result from the presence of an additional task.

Listening condition was contrast coded with two terms that compared 1) unmodified speech (-1) to the two levels of vocoding (0.5), and 2) 8-channel (-0.5) to 4-channel (0.5) vocoding. To assess the overall significance of this 3-level factor, we compared a model with both terms to one without using the χ^2 test of model comparison. Because an individual's success

in this task is likely influenced by their ability to simply match the grids, we also included average accuracy on the baseline grid-matching trials (centered) as a covariate in the grid-response model. The models used random intercepts by subject, as models with more complex random effect structures did not converge[2].

The speech-response model confirmed that accuracy decreased as listening condition became more difficult (Figure 2A). There was a significant effect of both levels of listening condition (Unmodified vs. Vocoded: B = -2.58, SE = 0.24, z = -10.85, $p < .001$; 8- vs. 4-channel: B = -1.56, SE = 0.09, z = -16.02, $p < .001$).

Our critical analysis concerned accuracy on the grid-matching task as a function of auditory difficulty (Figure 2B). As expected, baseline grid-matching ability significantly predicted dual-task performance (B = 0.45, SE = 0.09, z = 4.84, $p < 0.001$). We found an overall effect of listening condition when we compared this model to a reduced model that did not contain listening condition ($\chi^2(2) = 9.86$, $p = .007$). Examination of the two contrast codes suggested that performance decreased when individuals were matching grids while simultaneously hearing vocoded speech (M = 79.8) compared to unmodified speech (M = 83.1; B =-0.13, SE = .04, z = -2.91, $p = .004$), but no significant difference was found between the two levels of vocoding (8-channel vs. 4-channel; B = -0.03, SE = .07, z=-1.11, $p = .27$).. This confirms that the dual task was working as expected: grid-matching performance decreases when listeners must attend to both degraded speech and visual stimuli compared to unmodified speech

[2] Even models with only random intercepts by subjects and items failed to converge. We consulted with numerous mixed effects experts and were unable to resolve this. However, we ran separate models with only item intercepts found no diverging results for the dual task. For the pupillometry task, there was large variability amongst items which might suggest why our models with more complex random effects structures would not converge.

and visual stimuli. We also examined the pattern of RTs (see Supplement S1) and found largely similar results, with slower RT on grid matching trials for vocoded trials, but a small decrease in RT in 4-channel vs. 8-channel vocoding).

Pupillometry. Mean accuracy was 99.7% correct in the unmodified speech trials, 96.8% in the 8-channel vocoded trials, and 89.3% correct in the 4-channel vocoded trial (see Supplement S2 for an analysis of confusions). While participants largely recognized words correctly, vocoding increased difficulty. It is worth noting that because the pupil task was always done during the second visit, recognition of vocoded words is likely higher than in the dual task because of the previous visit's exposure to vocoding.

Figure 3 shows pupil dilation over time and as a function of listening condition. Pupil size peaked between 1000 and 1500 msecs, with later and higher peaks in the two degraded speech conditions. There did not appear to be substantial difference between the two vocoded conditions. It then fell to a trough at around 2000 and began climbing again before the appearance of the response screen.

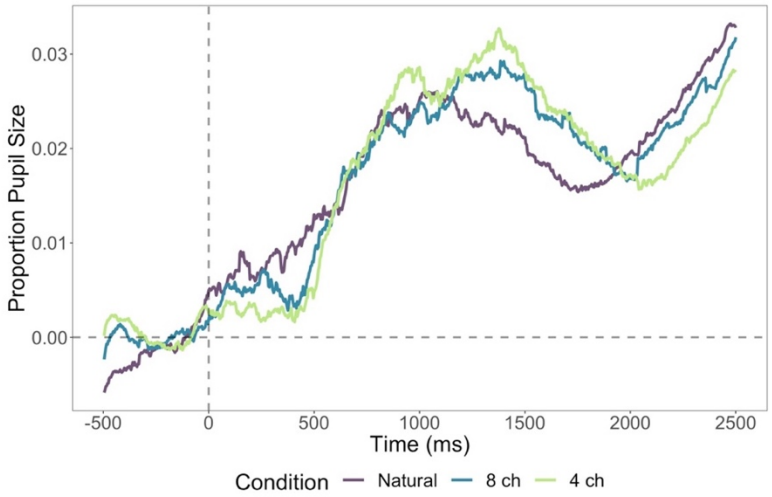

Figure 3. Proportion change in pupil size for each listening condition over time. The onset of the target word begins at 0 ms.

For analysis, maximum pupil size and time were used as the dependent variables in separate linear mixed effects models with listening condition as a predictor. Listening condition was contrast coded in the same way as in the dual task models. Random effects included random intercepts by subject as models with more complex random effects structures did not converge. Degrees of freedom were estimated using Satterthwaite's method as implemented by the lmerTest package (Kuznetsova et al., 2017) in R.

Listening condition was not a significant predictor for either maximum pupil size or time. This was shown in a likelihood ratio test that found no difference between these models and reduced models that did not contain listening condition as a factor (pupil size: $\chi^2(2) = 2.31$, $p = .31$, pupil timing: $\chi^2(2) = 0.84$, $p = .66$). This was confirmed in the individual coefficients (Unmodified vs. vocoded: B = 0.003, SE = 0.002, t(56) = 1.25, p =.22; 8-channel vs. 4-channel: B = 0.003, SE = 0.004, t(56) = 0.85, p = .39) or time (Unmodified vs. vocoded: B = 16.17, SE = 58.64, t(56) = 0.28, p = .78; 8-channel vs. 4-channel: B = -87.57, SE = 101.56, t(56) = -0.86, p = .39). Thus, while the descriptive results appear to support at least a difference between unmanipulated and manipulated speech, there was no statistical evidence for this.

Communality Analyses

One explanation for the absence of a significant effect of condition in the pupillometry is that pupil size may not solely reflect the difficulty of the task (e.g., the degree of vocoding), but may also reflect other processes like response planning or anticipation of feedback. These are likely to be related to the accuracy in that condition. We thus asked whether listening condition and accuracy each exert a unique effect on effort in each task (measured by grid-matching performance or pupil size), over and above the other factor.

To address this, we used a communality analysis. This is an older statistical approach that derives from hierarchical regression (Ray-Mukherjee et al., 2014), but can easily be implemented in a mixed effects framework. It is implemented by comparing a series of models whose fixed effects differ systematically. However, unlike stepwise selection approaches (which have been shown to lead to biased estimates and Type I errors: see Thompson, 1995; Whittingham et al., 2006), this is intended as a hypothesis driven approach to estimate the unique and shared variance among a set of predictors. This was conducted as a series of mixed models. In the first model, a single factor (e.g., listening condition) was entered in the model. The second model then added the other factor (accuracy). Critically, if the second model offered an improved fit, then second factor (accuracy) accounted unique variance over and above the first (listening condition). By reversing the order, we can then identify the unique variance associated with the listening condition.

Thus, we ran a series of models using a communality approach to investigate the degree to which a listener's accuracy in the task predicted pupil size and timing. Speech accuracy was the participants' mean accuracy in each listening condition, and listening condition was coded as before. Again, we used random intercepts by subject as more complicated random effects structures did not converge.

Tables 2 and 3 show the results. In the first-level model, speech accuracy (the only predictor) significantly predicted maximum pupil size (B = -0.06, SE = 0.03, t(58) = -2.07, p = .04), with more accurate subjects generally showing lower maximum pupil sizes. It did not predict maximum pupil time (B = -343.88, SE = 810.50, t(63) = -0.42, p = .67). However, in the second ste*p* of the model, listening condition did not predict any significant variance in pupil size. When we reversed the model, testing the effect of condition in the first step and accuracy in

Table 2. *Results from a communality analysis comparing speech accuracy and listening condition for maximum pupil size. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**				
1	Max pupil ~ Condition	Max pupil ~ Speech accuracy				
2	Max pupil ~ Condition + speech accuracy	Max pupil ~ Speech accuracy + Condition				
χ^2	2.48 (df=1, p=0.12)	0.58 (df=2, p=0.74)				
L2 Model						
Random effects		**Name**	**Variance**	**SD**		
Subject		Intercept	0.001	0.035		
Residual			0.0002	0.014		
Fixed Effects		**Estimate**	**SE**	**t**	**df**	***p***
Intercept		0.15	0.06	2.38	62.5	0.02
Listening Condition (unmodified vs. vocoded)		-0.002	0.004	-0.55	59.5	0.58
Listening Condition (8-channel vs. 4-channel)		-0.005	0.006	-0.75	59.3	0.46
Speech accuracy		-0.11	0.07	-1.55	61.6	0.13

Table 3. *Results from a communality analysis comparing speech accuracy and listening condition for maximum pupil time. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**				
1	Max time ~ Condition	Max time ~ Speech accuracy				
2	Max time ~ Condition + speech accuracy	Max time ~ Speech accuracy + Condition				
χ^2	2.164 (df=1, p=0.10)	3.29 (df=2, p=0.19)				
L2 Model						
Random Effects		**Name**	**Variance**	**SD**		
Subject		Intercept	143711	379.1		
Residual			196312	443.1		
Fixed Effects		**Estimate**	**SE**	**t**	**df**	***p***
Intercept		3662.17	1574.2	2.33	79.3	0.02
Listening Condition (unmodified vs. vocoded)		-104.07	94.95	-1.09	71.5	0.28
Listening Condition (8-channel vs. 4-channel)		-287.29	160.26	-1.29	71	0.07
Speech accuracy		-2642.72	1652.39	-1.6	79.1	0.11

the second, accuracy did not account for unique variance on the second step. As mentioned in the previous section, listening condition did not significantly predict pupil size in the first-level model on its own.

As a whole, this suggests that pupil size, in this task, largely reflects shared variance between accuracy and listening condition – neither had unique effect. However, the significant effect accuracy in the first level model (coupled with the lack of an effect of degradation),

suggests dilation is more strongly related to accuracy in this task. This is consistent with the idea that it partially reflects performance projection or anticipation of an error. Subjects (or conditions within subjects) with better accuracy show smaller pupil sizes, suggesting that the pupil size may have reflected subjects monitoring of their own performance.

In this case, the presence of feedback on each trial may have provided a cue to performance level and may have also created a situation in which subjects were dilating in anticipation of negative reinforcement. It is important to note though that in the second-level model when listening condition is added to the model containing speech accuracy for maximum pupil size, accuracy drops below significance (B = -0.11, SE = 0.07, t(61) = -1.55, p = .13). This suggests that even though condition is not a major factor, there may be shared variance with accuracy. The communality analysis for the pupillometry task suggests that the role of feedback might be obscuring the role of listening condition.

We next ran a similar analysis with the dual task data to confirm that listening condition (which was significant individually in the prior analysis) explains the majority of the variance above and beyond the role of speech accuracy. Here, the dependent variable was grid-matching performance in the dual-task trials. Speech accuracy was quantified as each participant's mean accuracy when responding to speech in each listening condition on the dual-task trials. Listening condition was contrast coded in the same way as described in the main dual task analysis. We also included baseline grid accuracy as a covariate. All models included random intercepts by subject as more complicated random effects structures did not converge.

Results of these analyses are reported in Table 4. First, speech accuracy did not account for any variance over and above listening condition alone. In contrast, in the reversed model, listening condition was significant over and above speech accuracy. This suggests that dual task

Table 4. *Results of a communality analysis comparing listening condition and speech accuracy for grid-matching performance on the dual-task trials. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**			
1	Grid response ~ Baseline + Condition	Grid response ~ Baseline + Speech accura			
2	Grid response ~ Baseline + Condition + speech accuracy	Grid response ~ Baseline + Speech accura + Condition			
χ^2	0.41 (df=1, p=0.5)	6.26 (df=2, p=0.04)			
L2 model					
Random Effects		**Name**	**Variance**	**SD**	
Subject		Intercept	0.034	0.19	
Fixed Effects		**Estimate**	**SE**	**z**	
Intercept		1.74	0.39	4.40	<0.
Baseline grid-matching accuracy		0.45	0.09	4.85	<0.
Listening Condition (unmodified vs. vocoded)		-0.16	0.06	-2.50	0
Listening Condition (8-channel vs. 4-channel)		-0.13	0.11	-1.23	0
Speech accuracy		-0.28	0.44	-0.64	0

performance is largely predicted by how hard the listening condition is, and not how successful participants are at speech perception.

These analyses suggest that the pupil and dual-task versions of the effort measures may be tapping distinct processes: dual-task is sensitive primarily to how difficult the speech is, while the pupil task seems to partially, but not completely, track accuracy.

Relationship between tasks

Next. we investigated the relationship between measures of listening effort. For this, we calculated difference scores for both the dual task and pupillometry as individual measures that can be correlated across tasks. We did not expect a correlation, given our unexpected finding that the two pupil measures were not predicted by listening condition. Nonetheless, these analyses were planned and had the potential to be informative in guiding our follow-up experiments.

An individual's dual task score was calculated as the difference between their accuracy on grid-matching dual task trials in unmodified speech and their accuracy on grid-matching dual

task trials in vocoded speech. Because we found no difference between 8- and 4-channel vocoding these were averaged together. Difference scores for the pupil task were calculated as the difference between the average pupil size or time on vocoded trials (again, 8- and 4-channel were averaged) and the same in the unmodified speech trials. Note this subtraction was reversed relative to the dual-task so that a positive score is indicative of increased effort in the vocoded conditions in both tasks (increased pupil size/later time and decreased accuracy in the dual task).

The dual-task difference score was moderately correlated to the pupil-size difference score (r = 0.39, p = .03; Figure 4A), but was not correlated to the pupil-time difference score (r = -0.01, p = .95; Figure 4B). This suggests that individuals with larger dual-task scores also show larger differences in pupil size. Given that pupil size seems to reflect more feedback response-monitoring than listening condition (vocoding level), this relationship is hard to interpret, but still suggests that the individuals who engage more effort are the same across tasks. This will be clarified as we refine the pupil task in Experiments 2 and 3.

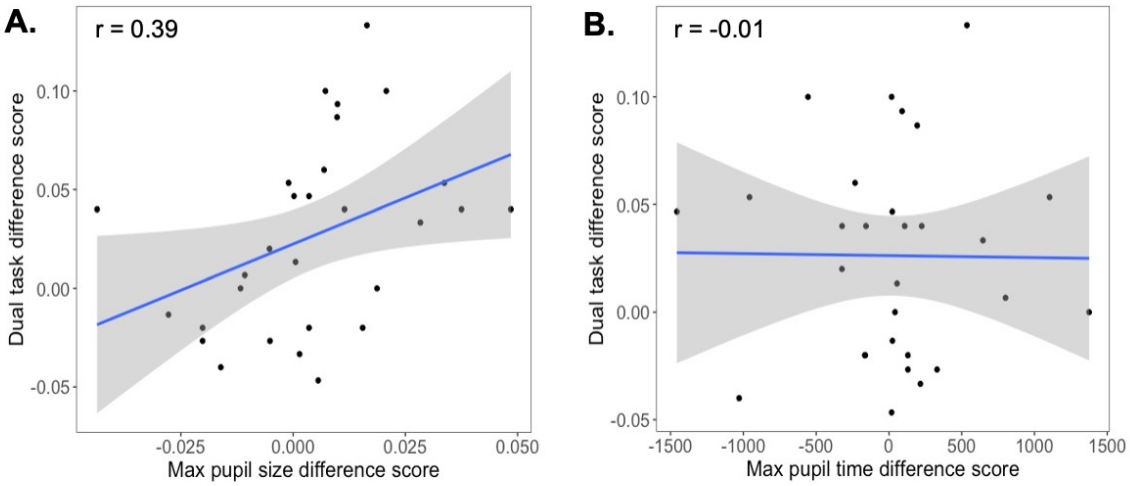

Figure 4. Relationship between dual task difference score and (A) maximum pupil size difference score and (B) maximum pupil time difference score

Relationship to Speech Perception Accuracy

Finally, we asked if effort helps achieve greater speech recognition accuracy. We ran additional mixed-effects models predicting speech accuracy for both the dual task and pupillometry. The dual task model predicted accuracy on each trial (1/0) on the speech-response dual-task trials as a function of listening condition and dual task difference score. We asked if effort (the difference score) uniquely predicted speech perception, accounting for the effect of condition. Table 5A shows the results. While listening condition remained significant ($p < .001$),

Table 5. *Results of the logistic mixed-effects models predicting speech accuracy in (A) the dual task, (B) the pupillometry task with pupil size as a predictor and (C) the pupillometry task with pupil timing as a predictor.*

A. Dual task difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.33	0.58	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Intercept	3.28	0.17	18.51	<0.001
Listening condition (unmodified vs. vocoding)	-2.58	0.24	-10.85	<0.001
Listening condition (8-channel vs. 4-channel)	-1.55	0.09	-16.02	<0.001
Dual task difference score	-0.40	2.26	-0.18	0.86
B. Pupil size difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.16	0.40	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Intercept	3.98	0.18	21.93	<0.001
Listening condition (unmodified vs. vocoding)	-2.07	0.21	-9.73	<0.001
Listening condition (8-channel vs. 4 channel)	-1.27	0.11	-11.64	<0.001
Max pupil size	-2.78	2.35	-1.18	0.24
C. Pupil time difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.15	0.39	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Intercept	3.84	0.13	28.78	<0.001
Listening condition (unmodified vs. vocoding)	-2.07	0.21	-9.73	<0.001
Listening condition (8-channel vs. 4-channel)	-1.31	0.11	-11.78	<0.001
Max pupil time (standardized)	-0.26	0.17	-1.49	0.13

the difference score did not explain any additional variation in accuracy beyond condition. The pupillometry models were similar, predicting speech recognition accuracy as a function of listening condition and either max pupil size or max pupil time as predictors. Similar to the dual-task model, neither maximum pupil size or time predicted recognition accuracy when accounting for listening condition (Table 5B and C). As a whole, this suggests that speech recognition accuracy is not related to effort (by either measure) over and above the larger effects of condition.

Experiment 1 Discussion

Experiment 1 confirms that our dual-task paradigm captures the additional effort required by the more challenging listening conditions. The pupillometry task, on the other hand, produced unexpected results. Accuracy, not listening condition, predicted maximum pupil size, though neither effect was uniquely significant in the communality analyses predicting effort.

We found mixed evidence when the tasks were related to each other. First, we found a small but moderate correlation between the dual-task and pupil-size difference scores. This suggests that they might be capturing something similar. However, at the same time, each task seemed responsive to different factors. Dual-task performance was only predicted by listening condition whereas pupillometry was related to perceived accuracy (though this was not significant when listening condition was included in the model).

These results must be qualified by the fact that pupil size was not sensitive to overall level of degradation. One possibility is that pupil size in this task may more be more sensitive to self-monitoring of performance because we provided feedback to participants on each trial. This is supported by the partial evidence that accuracy (but not condition) individually predicted pupil

size. We had included feedback in the first place to keep participants motivated, but it is possible that feedback was obscuring our results as it was encouraging greater error monitoring or anticipation of feedback. Given previous evidence for the influence of feedback on pupil response (Zekveld et al., 2019), we redesigned the task to remove this additional factor. Thus, we revised the pupillometry experiment to eliminate the feedback and ran a new group of subjects in Experiment 2. If this experiment successfully captured an effect of degradation on pupil size, we planned to run Experiment 3 to compare the new pupillometry task to the dual task of Experiment 1.

Experiment 2

Methods

Participants

Thirty-two participants were recruited from the University of Iowa's undergraduate Psychology subject pool. Participants received partial course credit for participating. Sample size was chosen to match Experiment 1.

Design

The same pupillometry task was used as in Experiment 1. The only change is that participants no longer received any feedback at the end of trials.

Stimuli

The same stimuli from Experiment 1 were used.

Procedure

Participants completed the updated pupillometry task in one visit to the lab. This task was identical to that used in Experiment 1 with the exception that the subjects received no feedback after their response. The procedure was otherwise the same as the second visit of Experiment 1.

Data Processing

The pupil response data were processed as described in Experiment 1. Participants were centrally fixating 65.4%, 66.3%, and 64.7% of the time during the unmodified, 8-channel, and 4-channel vocoding conditions, respectively. An average of 21.1% of samples were dropped because of blinks in the unmodified speech condition, 23% in the 8-channel condition, and 23% in the 4-channel condition.

Results

Mean accuracy in the pupillometry task was 99.7% correct in the unmodified speech trials, 93.2% in the 8-channel vocoded trials, and 78.3% correct in the 4-channel vocoded trials (see Supplement 2 for analysis of confusions). This is similar to Experiment 1, where the increasing difficulty in vocoding similarly affected accuracy.

Figure 5 shows pupil size as a function of time and condition. Results are much more consistent than Experiment 1 (c.f., Figure 3) with a clear effect of degradation (though likely no difference between the two levels of degradation). Further, the rise at the end of the timecourse was reduced, supporting the idea that this task variant is less sensitive to response demands.

We started by examining the effect of condition on effort, and then conducted communality analyses disentangling condition from accuracy. Finally, we asked if effort was

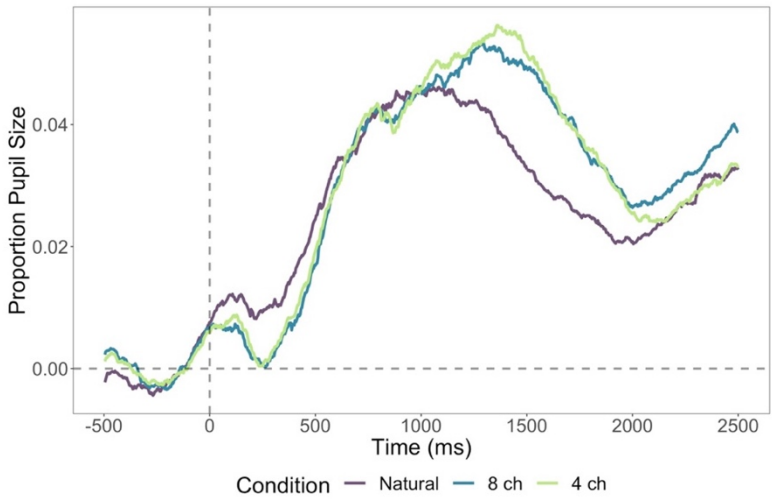

Figure 5. Proportion change in pupil size across the three listening conditions over time. The onset of the target word begins at 0 msec.

related to accuracy over and above condition.

Effects on Pupil response

For analyses, maximum pupil size and time were extracted following the same procedure as Experiment 1. Our first analysis used these in a linear mixed effects models with listening condition (contrast coded as in Experiment 1) as the fixed effect. Separate models were used for maximum pupil size and time. We again included random intercepts by subjects as more complicated random effects structures did not converge.

We started by comparing the full model with a reduced model that did not have listening condition as a fixed effect and found a significant difference between the two models ($\chi^2(2) = 11.83$, $p = .003$). This confirmed an overall significant effect of listening condition on pupil size. This was driven by a significant difference between unmanipulated speech and the vocoded conditions (B = 0.008, SE = 0.003, t(62) = 3.15, $p = .003$), but not between 8- and 4-channel vocoding (B = 0.007, SE = 0.005, t(62) = 1.64, $p = .11$). This suggests that when participants cannot track their performance through feedback, pupil size is larger under difficult listening

conditions, at least with respect to easy (unmodified) versus difficult (vocoded) conditions, if not at the level of distinguishing between levels of increasing difficulty (8-channel vs. 4-channel).

For peak pupil time, there was no overall significant effect of listening condition ($\chi^2(2) = 3.55$, $p = .17$). However, we found a trend towards a later peak pupil time in more difficult listening conditions (Figure 5), though this was not significant (Unmodified vs. vocoded: B = 92.15, SE = 49.12, t(57) = 1.88, $p = .06$). There was no significant difference in the timing of the peak pupil size between the two vocoding conditions (8-channel vs 4-channel: B = 5.79, SE = 84.47, t(58) = 0.07, $p = .95$).

Communality Analysis

We next conducted a communality analysis to ask whether listening condition now explains unique variance in pupil size and time above and beyond accuracy. Again, this used a series of linear mixed effects models that predicted maximum pupil size (Table 6) or time (Table 7). Models were run in two levels with the first level using only one of the two factors (e.g., listening condition) and the second level examining the unique effect of the other factor (accuracy). All models included random intercepts by subject.

Unlike Experiment 1, listening condition uniquely predicted maximum pupil size over and above accuracy (Table 6, left column). In contrast, in the reversed model (Right column), adding accuracy to the model with listening condition did not account for new variance. For timing, no combination of models was significantly different from each other. Thus, at least for maximum pupil, the pupillometry task without feedback shows a pattern that much more closely mirrors dual task performance.

Table 6. *Results from a communality analysis comparing speech accuracy and listening condition for maximum pupil size. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**			
1	Max pupil ~ Condition	Max pupil ~ Speech accuracy			
2	Max pupil ~ Condition + speech accuracy	Max pupil ~ Speech accuracy + Condition			
χ^2	3.14 (df=1, p=0.07)	10.50 (df=2, p=0.005)			
L2 Model					
Random Effects		**Name**	**Variance**	**SD**	
Subject		Intercept	0.003	0.05	
Residual			0.0003	0.02	
Fixed Effects		**Estimate**	**SE**	**t**	**df**	***p***
Intercept		0.015	0.37	0.42	72.6	0.68
Listening Condition (unmodified vs. vocoded)		0.014	0.004	3.26	63.9	0.002
Listening Condition (8-channel vs. 4-channel)		0.018	0.007	2.39	63.8	0.02
Speech accuracy		0.07	0.04	1.75	65.4	0.08

Table 7. *Results from a communality analysis comparing speech accuracy and listening condition for maximum pupil time. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**			
1	Max time ~ Condition	Max time ~ Speech accuracy			
2	Max time ~ Condition + speech accuracy	Max time ~ Speech accuracy + Condition			
χ^2	0.44 (df=1, p=0.51)	1.63 (df=2, p=0.44)			
L2 Model					
Random Effects		**Name**	**Variance**	**SD**	
Subject		Intercept	57946	240.7	
Residual			105171	324.3	
Fixed Effects		**Estimate**	**SE**	**t**	**df**	***p***
Intercept		1549.29	590.99	2.62	83.1	0.01
Listening Condition (unmodified vs. vocoded)		53.09	76.87	0.69	72.5	0.49
Listening Condition (8-channel vs. 4-channel)		-58.55	128.91	-0.45	71.1	0.65
Speech accuracy		-429.67	651.71	-0.56	83	0.51

Relationship to Speech Perception Accuracy

Finally, we asked which factors predicted speech perception accuracy. For this we used logistic mixed effects models that predicted response (1/0) in the new pupillometry task from

listening condition and either maximum pupil size or time. Random intercepts by subject were the only random effect as more complex random effects structures failed to converge.

Listening condition significantly predicted accuracy, with significant effects for the unmanipulated vs. vocoding contrast and for the 8- vs. 4-channel vocoding (see Table 8A and B). However, neither maximum pupil size (B = 1.55, SE = 1.44, z = 1.08, p = .28), nor pupil time (B = -0.11, SE = 0.14, z = -0.78, p = .44) predicted accuracy. Thus, even with an improved pupil task that is sensitive to the level of degradation, we still did not find any evidence that effort (by either pupil size or timing) is related to accuracy.

Table 8. *Results of logistic mixed-effects models predicting speech accuracy in the pupillometry task using listening condition and A) maximum pupil size or B) maximum pupil time.*

A.) Maximum pupil size difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.23	0.48	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Intercept	3.19	0.17	18.41	<0.001
Listening condition (unmodified vs. vocoding)	-2.58	0.19	-13.29	<0.001
Listening condition (8-channel vs. 4 channel)	-1.39	0.08	-18.09	<0.001
Max pupil size difference score	1.55	1.44	1.08	0.28
B.) Maximum pupil time difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.24	0.49	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Listening condition (unmodified vs. vocoding)	-2.51	0.20	-12.38	<0.001
Listening condition (8-channel vs. 4-channel)	-1.38	0.08	-16.98	<0.001
Max pupil time difference score (standardized)	-0.11	0.14	-0.78	0.44

Experiment 2 Discussion

When we removed feedback from the pupillometry task, difficulty of the task (listening condition) predicted maximum pupil size. The contrast with Experiment 1 suggests first, that pupillometry can be highly sensitive to task factors (like the availability of feedback) that may

not be related to how hard the participant finds the task. Second and more importantly, this version was sensitive to listening condition, unlike in Experiment 1, where we found that an individual's accuracy was more predictive of their pupil dilation. Despite this, we still did not find any evidence that effort, as indexed by either pupil size or pupil timing, predict how well listeners will perform at speech perception. With this improved a pupillometry task that reflects listening condition (vocoding level), Experiment 3 returned to our original goal of assessing the relationship between measures of effort.

Experiment 3

Experiment 3 ran an additional group of participants on both tasks to again ask if there is any relationship between performance on the two tasks of listening effort.

Methods

Participants

Fifty-two subjects were recruited from the University of Iowa community. Participants received either partial course credit or monetary reimbursement for participating. Twenty-two subjects are excluded (two did not finish the tasks during their visits, one had low accuracy, and nineteen were lost due to technical issues with the eye tracker[3]), leaving 30 full data sets. Power was based on Experiment 1.

Stimuli

The same stimuli were used for both tasks as in the previous experiments.

[3] This was due to a system upgrade that disrupted the ability to save data over the network between the host and operator computers.

Design

The dual task from Experiment 1 was used, and the pupillometry task from Experiment 2.

Procedure

Across two visits to the lab, participants completed the dual task (on their first visit) and the pupillometry task (one week later). The procedures for both tasks were otherwise unchanged.

Data Processing

The dual task and pupil response data were processed as described in Experiment 1. For the pupillometry task, participants were centrally fixating during 67.2%, 66.8%, and 65.6% of the samples during the unmodified, 8-channel, and 4-channel vocoding conditions, respectively. An average of 14.0% of samples were dropped because of blinks in the unmodified condition, 15.7% in the 8-channel condition, and 18.1% in the 4-channel condition.

Results

We first examined the effect of listening condition for each task individually to document that degradation had the predicted effect on effort. We then followed up on this with a communality analysis to ask how listening condition and accuracy independently predict effort. Next, we turned to our primary question to determine if the two tasks are related. Finally, we related both effort and condition to accuracy to ask if effort improves speech perception.

Effect of Condition on Effort

Dual Task. Mean accuracy for all the trial types in the dual task is reported in Table 9 and Figure 6. As expected, performance decreased with increasingly degraded speech: from 99% accuracy with unmanipulated speech, to 73% accuracy in 4-channel speech. The same pattern is seen for the grid-matching trials, although the decrease is smaller: baseline accuracy with no speech presented is about 83%, which drops to 80% accuracy when simultaneously presented with 4-channel speech.

Table 9. *Mean accuracy by trial type and listening condition in the dual task.*

		Baseline	Unmanipulated Speech	8-channel	4-channel
Single task	Grid response	83.0%			
Dual task	Speech response		99.8%	89.9%	73.4%
	Grid response		82.5%	80.6%	80.2%

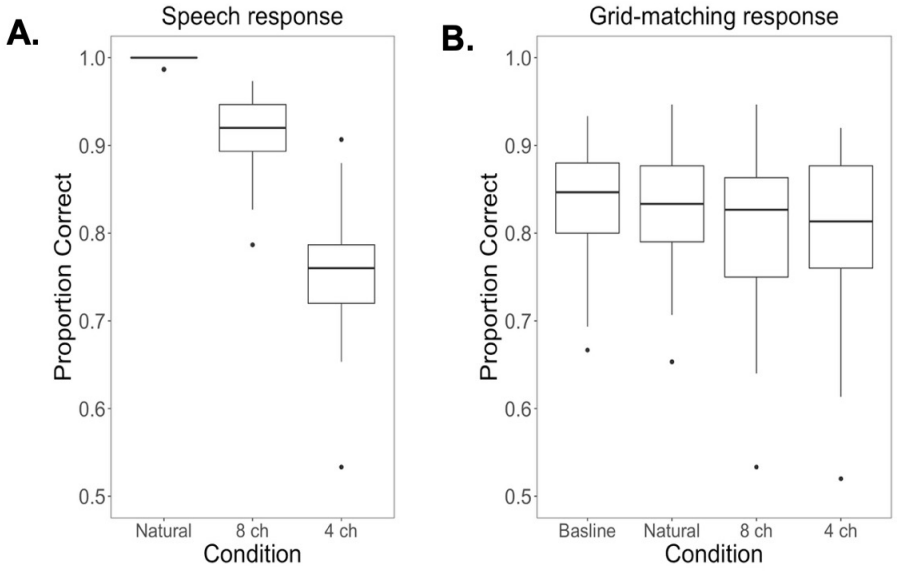

Figure 6. Proportion correct responses in each listening condition of the dual task for (A) the speech-response trials and (B) the grid-response trials.

To support this statistically, we ran separate mixed effects logistic regressions predicting response (1/0) in the speech-response trials and the grid-matching trials respectively. Listening condition was contrast coded as previously described for both models, and baseline grid-matching accuracy was centered and included in the grid-matching model. Random intercepts by subject were included as any further addition to the random effects structure did not converge.

Listening condition was significant at both levels of the speech-response model (Unmodified vs. vocoded: B = -3.08, SE = 0.27, z = -11.22, $p < .001$; 8-channel vs. 4-channel: B = -1.27, SE = 0.08, z = -15.26, $p < .001$). This confirms that performance was significantly worse as speech became more degraded.

The critical analysis was the grid-matching model. As with Experiment 1, baseline grid-matching ability significantly predicts grid-matching during the dual-task trials (B = 0.74, SE = 0.10, z = 7.14, $p < .001$). Crucially, listening condition as a whole was significant ($\chi^2(2) = 5.65$, $p = .05$). This was driven by a significant difference between unmanipulated and vocoded speech (B = -0.1, SE = 0.04, z = -2.33, $p = .02$), suggesting that as speech perception became difficult, grid-matching performance suffered. This confirms that the dual-task was sensitive to resource-limited effort deployed for the task. There was no significant difference between the two levels of vocoding (8-channel vs. 4-channel: B = -0.03, SE = 0.07, z = -0.4, $p = 0.69$), suggesting that the task was not sensitive to distinctions among levels of degradation (as we have seen in both of the prior experiments). Analysis of RTs showed no significant effects (Supplement S1).

Pupillometry. Mean accuracy in each listening condition for the pupillometry task was similar to the speech-response trials of the dual task: 99.6% for unmodified speech, 93.2% for 8-channel vocoding, and 82.9% for 4-channel vocoding. Figure 7 plots the timecourse of pupil size relative to baseline in each listening condition.

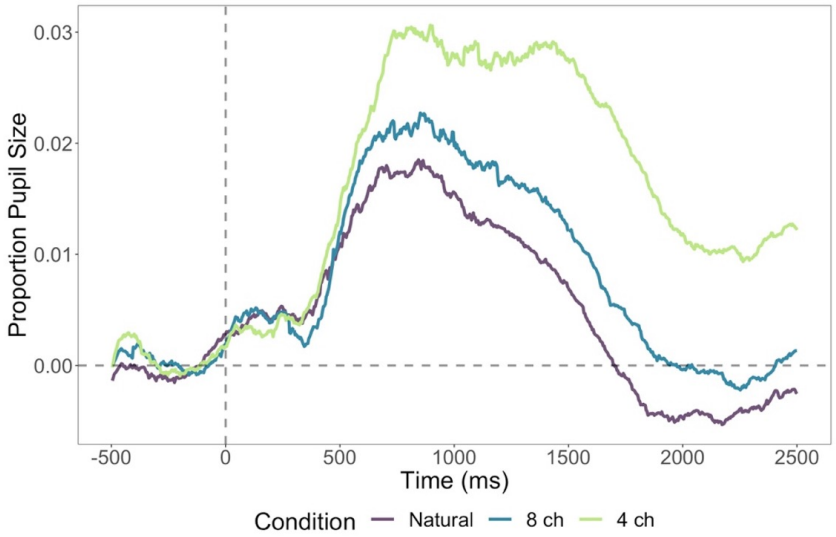

Figure 7. Proportion change in pupil size during each listening condition. The onset of the target word began at 0 msec.

To investigate the relationship between pupil response and listening condition, we ran two linear mixed effects models predicting either maximum pupil size or time. Listening condition was contrast coded identically to previous models. Random intercepts by subjects were included.

Listening condition as a whole significantly predicted maximum pupil size ($\chi^2(2) = 17.09$, df = 2, $p < .001$). Individual contrasts showed significant effects of both the Unmodified vs. vocoded contrast (B = 0.007, SE = 0.002, t(55) = 3.26, $p = .002$) and the 8-channel vs. 4-channel contrast (B = 0.01, SE = 0.004, t(55) = 2.94, $p = .005$). Maximum pupil time showed no overall effect ($\chi^2(2) = 3.03$, df = 2, $p = 0.22$) and neither contrast was significant (Unmodified vs. vocoded: B = 41.64, SE = 51.72, t(56) = 0.81, $p = .42$; 8-channel vs. 4-channel: B = 134.06, SE = 87.75, t(55) = 1.52, $p = .13$). This suggests that as the listening condition became more difficult through increased degradation, maximum pupil size increased. However, increased degradation did not affect the timing of the peak pupil size.

Communality Analyses

We again asked if speech recognition accuracy and listening condition uniquely explain variance in each task. The goal of this analysis was to confirm that our two measures of effort are driven by similar underlying factors, mainly listening condition, replicating the prior claims of Experiment 1 for the dual task, and Experiment 2 for pupillometry. Models were run in two levels with the first level using only one of the two factors (e.g., listening condition) and the second level examining the unique effect of the other factor (accuracy). All models included random intercepts by subject, as more complex random effects structures did not converge.

For the dual task, the dependent variable was grid-matching performance in the dual-task trials. Participants' mean accuracy in each listening condition was again used as a measure of speech accuracy. Listening condition was contrast coded in the same way as described in Experiment 1. We also included baseline grid accuracy as a covariate. Results are in reported in Table 10.

Table 10. *Results of a communality analysis comparing listening condition and speech accuracy for grid-matching performance on the dual-task trials. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**			
1	Grid response ~ Baseline + Condition	Grid response ~ Baseline + Speech accuracy			
2	Grid response ~ Baseline + Condition + speech accuracy	Grid response ~ Baseline + Speech accuracy + Condition			
χ^2	1.48 (df=1, p=0.22)	5.62 (df=2, p=0.06)			
L2 model					
Random Effects		**Name**	**Variance**	**SD**	
Subject		Intercept	0.052	0.23	
Fixed Effects		**Estimate**	**SE**	**Z**	***p***
Intercept		0.83	0.36	2.32	0.02
Baseline grid-matching accuracy		0.76	0.09	7.81	<0.001
Listening Condition (unmodified vs. vocoded)		-0.09	0.04	-2.13	0.03
Listening Condition (8-channel vs. 4-channel)		-0.03	0.07	-0.41	0.68
Speech accuracy		0.76	0.39	1.67	0.09

Speech accuracy did not explain any variance over and above listening condition (p = .22), nor did listening condition explain variance over and above speech accuracy (p = .06). However, the coefficient for the unmodified vs. vocoded contrast in the second-level model was significant (p = .03) (though the contrast between 4- and 8-channel vocoding was not). This suggests that 4- vs. 8- channel contrast may be what brought down the overall significance (which includes both contrasts). Thus, there is some evidence that, as in Experiment 1, the dual task was sensitive to condition but not speech recognition accuracy.

For the pupillometry task, we again ran parallel analyses for maximum pupil size and time. The pupil size model comparison is reported in Table 11. We found clear evidence that listening condition uniquely explains variance over and above accuracy alone. When this was reversed (speech accuracy added to the condition-only model), there was no difference between models, suggesting that listening condition is the more important of the two factors. The analysis for pupil time is reported in Table 12. Unlike pupil size, neither listening condition nor accuracy uniquely explained variance in the timing of the pupil response.

Table 11. *Results from a communality analysis comparing speech accuracy and listening condition for maximum pupil size. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**
1	Max size ~ Condition	Max size ~ Speech accuracy
2	Max size ~ Condition + speech accuracy	Max size ~ Speech accuracy + Condition
χ^2	0.18 (df=1, p=.67)	10.12 (df=2, p=.006)

L2 Model

Random Effects	**Name**	**Variance**	**SD**		
Subject	Intercept	0.0004	0.02		
Residual		0.0002	0.01		
Fixed Effects	**Estimate**	**SE**	**t**	**df**	***p***
Intercept	0.03	0.02	1.25	77.1	.22
Listening Condition (unmodified vs. vocoded)	0.008	0.003	2.71	64.2	.009
Listening Condition (8-channel vs. 4-channel)	0.01	0.004	2.66	60.9	.01
Speech accuracy	0.01	0.02	0.42	74.9	.67

Table 12. *Results from a communality analysis comparing speech accuracy and listening condition for maximum pupil time. The second-level model is summarized fully below the model comparison.*

	Accuracy | Condition	**Condition | Accuracy**
1	Max time ~ Condition	Max time ~ Speech accuracy
2	Max time ~ Condition + speech accuracy	Max time ~ Speech accuracy + Condition
χ^2	0.46 (df=1, p=.49)	1.07 (df=2, p=.58)

L2 Model					
Random Effects	**Name**	**Variance**	**SD**		
Subject	Intercept	33768	183.8		
Residual		113396	336.7		
Fixed Effects	**Estimate**	**SE**	**t**	**df**	***p***
Intercept	1237.45	439.92	2.81	76.6	.006
Listening Condition (unmodified vs. vocoded)	17.14	63.59	0.37	73.9	.79
Listening Condition (8-channel vs. 4-channel)	101.90	100.27	1.016	67.4	.32
Speech accuracy	-316.07	475.39	-0.67	77.3	.51

These analyses suggest that both the dual task and pupillometry task pattern in the same direction. That is, listening condition is an important factor for explaining variance in both tasks. This suggests that both tasks are now sensitive to how difficult speech perception is.

Relationship Between Tasks

Finally, we investigated the relationship between tasks. As in Experiment 1, we calculated difference scores for each task as an index of an individual's effort. The dual task difference score was the difference between grid-matching performance on unmanipulated trials and vocoded trials. The two pupillometry differences scores were the difference between the average pupil (either size or time) on the vocoded and unmanipulated trials.

Figure 8 shows the relationship between dual task difference score against and either the maximum pupil size difference score (8A) or their maximum pupil time difference score (8B). Despite having much better matched tasks (than Experiment 1), we found no correlation with either the maximum pupil size (r = 0.19, p = .34) nor maximum pupil time difference scores (r =

-0.11, p = .59). This suggests that even when dual tasks and pupillometry are matched for explicit task and stimuli, there is no evidence for a relationship between them.

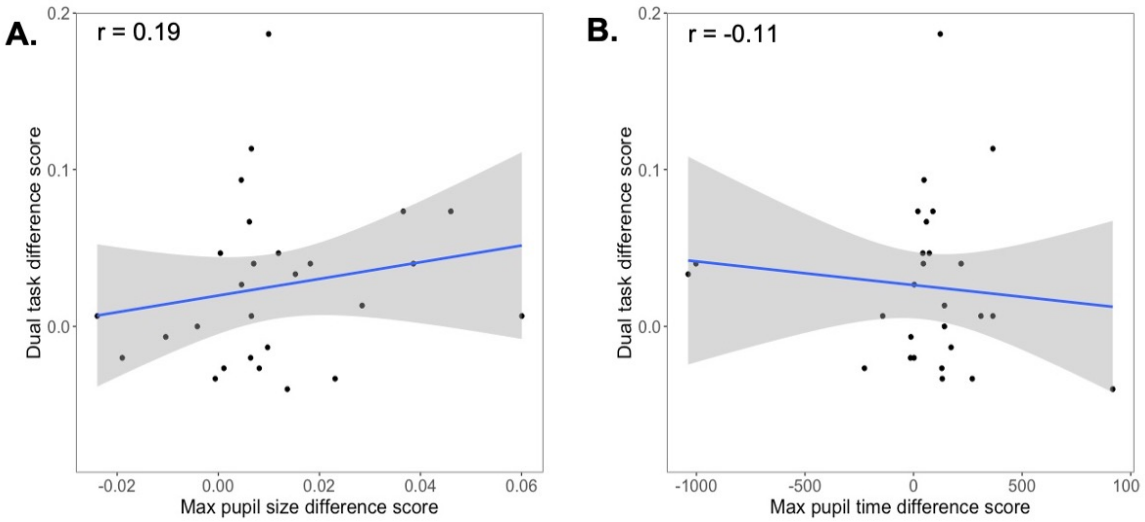

Figure 8. Relationship between the dual task difference score and (A) pupil size difference score or (B) pupil time difference score.

Relationship to Speech Perception Accuracy

Finally, we return to our question of whether effort is related to speech perception performance. We ran three additional logistic mixed effects models for each of our effort metrics. For the dual task, speech response (1/0) was predicted by a model containing listening condition (contrast coded as in Experiment 1) and the dual task difference score described in the previous section. For the pupillometry task, speech response (1/0) was predicted by a model containing listening condition (contrast coded in the same way) and either the maximum pupil size or time difference score. All three models included random intercepts by subject. More complex random effects structures did not converge.

Table 13 presents the results for all three models. In all cases, listening condition significantly predicted speech response both when comparing unmodified to vocoded speech

(p<.001) and when comparing 8-channel to 4-channel ($p < .001$). At no point did any of the measures of effort predict speech perception accuracy (Dual task difference score: B = 0.33, SE = 1.94, z = 0.17, $p = .86$; Pupil size difference score: B = -1.25, SE = 3.25, z = -0.38, $p = .70$; Pupil time difference score: B = -0.13, SE = 0.11, z = -1.21, $p = .23$). Consistent with previous

Table 13. *Results of the logistic mixed-effects models predicting speech accuracy in (A) the dual task, (B) the pupillometry task with pupil size as a predictor and (C) the pupillometry task with pupil timing as a predictor.*

A. Dual task difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.37	0.61	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Intercept	3.24	0.18	18.01	<0.001
Listening condition (unmodified vs. vocoding)	-3.08	0.27	-11.23	<0.001
Listening condition (8-channel vs. 4-channel)	-1.27	0.08	-15.26	<0.001
Dual task difference score	0.33	1.94	0.17	0.86
B. Pupil size difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.38	0.62	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Intercept	3.54	0.19	17.77	<0.001
Listening condition (unmodified vs. vocoding)	-2.46	0.18	-12.98	<0.001
Listening condition (8-channel vs. 4 channel)	-1.13	0.09	-12.13	<0.001
Max pupil size	-1.25	3.25	-0.38	0.70
C. Pupil time difference score				
Random Effects	**Name**	**Variance**	**SD**	
Subject	Intercept	0.38	0.62	
Fixed Effects	**Estimate**	**SE**	**Z**	***p***
Intercept	3.49	0.15	23.14	<0.001
Listening condition (unmodified vs. vocoding)	-2.46	0.19	-12.99	<0.001
Listening condition (8-channel vs. 4-channel)	-1.12	0.09	-12.63	<0.001
Max pupil time (standardized)	-0.13	0.11	-1.21	0.23

evidence that speech intelligibility is separable from engaged effort (e.g., Winn & Teece, 2020), we found no evidence that listening effort predicts speech perception accuracy.

Experiment 3 Discussion

Experiment 3 found that listening condition uniquely predicted effort in both the dual task and pupillometry task (at least for pupil size, if not for the timing of the pupil response). Despite this, we found no relationship between individual performance on the dual task and pupil response. We also found no evidence that listening effort, as indexed by our three difference scores, predicts speech perception accuracy. While effort may play an important role in speech perception under difficult condition, we found no evidence that it actually improves performance.

General Discussion

Three experiments investigated the relationship between cognitive and physiological measures of listening effort. Our goal was to compare behavior in closely matched listening effort tasks to discover if these measures are related within individuals and to examine whether this behavior predicted speech perception accuracy.

Evaluating the Listening Effort Tasks Individually

We did not use "off the shelf" listening effort tasks in this study for several reasons. First, it is important to note that this project was conducted in preparation for a larger longitudinal project examining outcomes in cochlear implant users. Second, with the immediate goal of relating these tasks to each other, it was important that tasks had similar demand characteristics. Third, in this larger project we would eventually relate these tasks to data from the Visual World

Paradigm (a central goal of the larger project). Thus, both tasks were modeled as a 4AFC single-word recognition task with a mouse-click response. Third and most importantly, we sought tasks that could not only measure group effects (e.g., a difference between conditions), but were also suitable for individual differences, so they could be related to each other (and to the Visual World Paradigm in our larger project).

Both tasks largely showed the expected patterns of group differences. In all three experiments, listeners showed evidence of increased effort (slower RTs on dual task trials, increased pupil dilation) when speech was vocoded vs. when it was not. This suggests that at the broadest level, these tasks do capture something about effort or difficulty. However, there were a number of factors that arose with each task that are worth consideration.

Dual Task Procedure. One concern with the validity of the tasks was the little evidence of changes in effort between 4- and 8-channel vocoding in the dual-task experiments. This was somewhat surprising as 4-channel vocoded speech was clearly harder (as indicated by accuracy in the speech-response trials). We suspect that this was because with 8-channel vocoding, most listeners exerted a reasonable amount of effort, however with 4-channel vocoding some of them put in enhanced effort while others "gave up" (c.f. Wu et al., 2016). This could explain the lack of difference, but it also could explain the unexpectedly faster RT in 4-channel vocoded speech in Experiment 1—when listeners "give up", they respond very quickly because they are guessing. This explanation is also consistent with the fact that such a difference was observed in the pupillometry task in Experiment 3: without the additional cognitive load imposed by the dual task people were less likely to give up.

Pupillometry. Experiment 1 also raised the possibility that pupillometry does not solely reflect engaged effort for the task at hand. The use of response feedback in Experiment 1 may

have encouraged a processing strategy that was reliant on progress monitoring. This was seen in the hierarchical regression where the objective level of degradation in the speech did not predict pupil dilation over and above accuracy. By removing feedback in Experiments 2 and 3, we were able to more accurately measure effort that individuals were engaging to compensate for the more difficult listening conditions and potentially avoid a situation where participants adjust their engaged effort because of how well they believe they are performing. Without feedback, this task showed effects of condition (including 4- vs. 8-channel vocoding) over and above accuracy. This suggests that the removal of feedback may have reduced the role of this kind of performance monitoring. Note that we are not arguing that listeners *only* engage such monitoring when there is feedback – indeed, even without feedback, subjects often engage in error-monitoring in difficult tasks (Ullsperger et al., 2010). However, this pattern of results is consistent with the idea that eliminating the feedback caused listeners to engage in less of this.

Several studies have used pupillometry to investigate the role of reward (Koelewijn et al., 2018), attention (Koelewijn et al., 2014, 2015, 2017), and feedback (Zekveld et al., 2019) on listening effort, suggesting that under certain task demands pupil size can be reflective of all these different factors. Through task instruction and design, experimenters can manipulate the factors to which pupil size is sensitive and caution should be taken to ensure the task is not inadvertently affecting participants' response. We further caution researchers and clinicians to avoid thinking of these types of measures as simply "off the shelf" measures of effort – it is not difficult to customize either a pupillometry or dual task measure to match the primary speech perception measure (similar to how we have designed the two tasks here to match). This may provide the important ability achieve complementary measures of how well a listener is doing,

and how much effort it takes them to get there, with one particular task with and a particular set of stimuli.

One of the benefits of using pupillometry is that it can assess effort that is occurring alongside speech perception, rather than inferring it from a secondary task as in the dual task. It is notable that our pupillometry task was sensitive to differences listening condition for words in isolation. Previous attempts at this were unsuccessful: Lau et al. (2019) found no difference in pupil size when comparing across different SNRs for words in isolation, but did find a difference when using sentences. We adapted the Visual World Paradigm to keep our task largely driven by recognition, while Lau et al. used a listen-and-repeat task that required speech production. Our task required listeners to map speech to meaning and is arguably a more naturalistic measure of effort during word recognition as it does not require additional resources to plan an utterance. It is possible that the cost of planning a production in Lau et al. (2019) pushed participants towards a ceiling of pupil dilation. We also scaled pupil response to each individual's dynamic range which allows us to account for variation in the reactiveness of participants' pupil. This may have proved important for picking up on smaller changes in pupil size that result from the shorter stimulus duration of a word compared to a sentence.

Importantly, only pupil size—but not pupil timing--was related to listening condition. This contrasts with previous work has found a relationship between the slope of the pupil response and auditory processing demands (Winn et al., 2015). It is possible that this is the fault of our method for extracting pupil timing, as we used an average time window around the peak pupil size and not the slope of the pupil response. It is also possible that words in isolation do not provide enough time to respond to the stimulus compared to listening to a degraded sentence, as in Winn et al. (2015). It could be necessary for listeners to be exposed to longer stimuli for a

pupil timing metric to be meaningful. Conversely the time of the peak may reflect not the effort of speech perception per se, but rather the effort engaged in sentence comprehension (with the poorer input). The timing of the pupil response could provide interesting insight into effortful processing, but it needs more refinement—particularly tying it to speech perception rather than other constructs—before being used to index listening effort.

Other Measures. Future studies seeking to assess listening effort should take care when considering their choice of measure. We did not include a subjective measure of effort in the present study, but there is evidence of a relationship between self-reported feelings of effort and objective measures of listening effort (Koelewijn et al., 2014; Picou et al., 2011; Picou & Ricketts, 2018). Our goal was to relate objective measures of effort, but it could be of interest to gauge a listener's self-awareness of their effort, especially for its fatiguing side effects. Studies should identify what aspects of effort they want to evaluate and carefully design measures of effort accordingly.

What makes listening difficult?

Throughout this manuscript we have used listening condition (unmodified vs. vocoded speech) as a proxy for the inherent difficulty of the stimulus. This is clearly an oversimplification, as part of the difficulty of identifying a word is in selecting it from amongst viable competitors—consequently the contents of the competitor set also matter (e.g., Luce & Pisoni, 1998). As the same set of items were used at all three levels of degradation, this is not likely to be a pure effect of lexical factors, but it may be an interaction of lexical and acoustic factors. Indeed, studies using reaction times and intelligibility (Clopper et al., 2006), and VWP

measures (Ben-David et al., 2011) have shown interactions between acoustic degradation and neighborhood density.

It might also be argued that the presence of the competitors in the response set (on pupil trials, or on speech trials in the dual-task paradigm) may enhance this effect. However, this is unlikely to be the case for two reasons. First, in the pupillometry task, the response options were not available until after the pupil measurement (unlike the VWP there was no prescan and there was even a delay until the appearance of the response options). Thus, at the time at which the pupil was measured (and at which effort was employed), this was functionally an open-set task and listeners would not have known what kind of trial it was. Second, in the dual-task on the vast majority of the trials, the lexical choices were never presented as the task was to do the visual matching task.

Even so, we would not rule out an interaction of the level of degradation and the competitor set in the mental lexicon. We do not draw a strict dichotomy between lexical and auditory processing (e.g., McMurray et al., 2002). Nonetheless, such an interaction is one that should be equally operative in both tasks (when relating them to each other), is equally relevant for accuracy, and would not be expected to change within an individual (e.g., if using these measures as individual differences measures).

Relating Dual Task and Pupil Measures

The group level effects suggest that each task individually shows the expected response to listening condition. However, when we examined individual differences a distinct story emerged. In Experiment 1 there was a moderate relationship between dual-task and pupil measures. However, recall that the pupil measure was not sensitive to listening condition at all,

and only to accuracy (while the converse was true for the dual-task). This undercuts the case that these tasks are related. In Experiment 3, when the tasks were now much better aligned, we found little to no evidence for a relationship between the dual task and pupillometry measures we employed.

These findings fit into a growing body of work that suggests listening effort is a complex construct which has yet to be well characterized by any one measure. Previous work comparing cognitive (dual-task) and physiological (pupillometry, EEG, skin conductance) measures of effort have found little relationship between tasks (Alhanbali et al., 2019; Strand et al., 2018), despite the fact that these distinct tasks can provide converging evidence for similar effects (e.g., Brown et al., 2020). We designed measures that were carefully matched on stimuli and task to better compare the engagement of effort in response to degraded speech. Moreover, by Experiment 3, both tasks showed largely the same profile of uniquely predictive factors (listening condition but not accuracy). However, even with these additional considerations in place, we still did not find a relationship between dual-task performance and pupil response.

This suggests that these different tasks are sensitive to different facets of listening effort and that the effort engaged for speech perception cannot be distilled into a singular mechanism. The FUEL (Pichora-Fuller et al., 2016) highlights the multidimensional nature of listening effort by emphasizing that effortful listening is subject to constraints on both an individual's general cognitive ability (e.g., resources) and motivation. The specifics of which aspects of cognitive ability are important for effort also remain unclear, leaving a broad variety of factors that can influence the engagement of effort. For example, effort likely reflects both the amount of resources available to a person, and also their skill at using them (e.g., differences in cognitive control or working memory), or people could vary in how many "resources" they need for a

given task. The dual task measures the cost of attending to two simultaneous tasks compared to each task on its own; thus, it may be more sensitive to resource availability or allocation. In contrast, the pupil response is directly tied to the responsiveness of the autonomic nervous system and may be more directly related to arousal-based processes. Thus, it is likely that these tasks are sensitive to different aspects of the underlying framework of listening effort, with pupillometry being more sensitive to attention and arousal, and dual task methods more sensitive to cognitive capacity.

What is the functional role of effort during speech perception?

Our results also highlight the need to further investigate the role of listening effort during speech perception in difficult listening situations. We found no evidence that effort predicts speech perception accuracy over and above listening condition. That is, engaging more effort did not mean an individual would be more successful recognizing words. This is consistent with several previous studies that find that intelligibility is separable from effort (Mackersie & Cones, 2011; Sarampalis et al., 2009; Winn & Teece, 2020). It is assumed that extra cognitive resources are engaged to compensate for challenging listening situations, but this extra effort seems to hit a ceiling of effectiveness for a listener's speech accuracy. At a certain point, putting in more effort is simply not enough to achieve improved accuracy.

Another possibility is that other various top-down strategies to compensate for poor input quality obscure the role of effort. For instance, word frequency effects could mislead a listener to recognizing a high-frequency word in noise over its correct but low-frequency competitor (e.g., mishearing *laud* as *loud*). Here, allocation of effort could actually enhance these cognitive biases.

We also need to take into account the possibility that accuracy reflects not only the contributions of effort, but also variation in (pre-effortful) ability. Listeners who are poor at perceiving vocoded speech automatically, may have lower effort, but also lower accuracy (essentially canceling out the benefit of effortful processing). At this point the role of effort for successful speech recognition remains an open question and there are many avenues of future inquiry to pin down the precise nature of listening effort. However, our results make it clear that there is not a large and obvious linkage between them.

Conclusions

Listening effort is an important construct for understanding speech perception in adverse listening conditions. However, much more work is required to understand the methods used to tap into engaged effort. It is becoming clear that different paradigms that all purport to measure listening effort are not strongly related to one another, raising the question of exactly what each paradigm is actually measuring. While it is hard to deny the importance of this construct for understanding the subjective experience of many hearing impaired listeners, our work suggests that translating this to a mechanistic and measurable construct may take significantly more work.

Acknowledgements

The authors would like to thank Kristin Rooff and Jamie Klein-Packard for assistance with data collection; Francis Smith for helpful comments on these ideas; and Jason Geller for consultation on pupillometry and statistical methods.

References

Ahern, S., & Beatty, J. (1979). Pupillary Responses During Information Processing Vary with Scholastic Aptitude Test Scores. *Science*, *205*(4412), 1289–1292.

Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of Listening Effort Are Multidimensional. *Ear and Hearing*, *40*(5), 1084–1097. https://doi.org/10.1097/AUD.0000000000000697

Ayasse, N. D., Lash, A., & Wingfield, A. (2017). Effort Not Speed Characterizes Comprehension of Spoken Sentences by Older Adults with Mild Hearing Impairment. *Frontiers in Aging Neuroscience*, *8*(329), 1–12. https://doi.org/10.3389/fnagi.2016.00329

Baldock, J., Kapadia, S., & van Steenbrugge, W. (2019). The task-evoked pupil response in divided auditory attention tasks. *Journal of the American Academy of Audiology*, *30*(4), 264–272. https://doi.org/10.3766/jaaa.17060

Ben-David, B. M., Chambers, C. G., Daneman, M., Pichora-Fuller, M. K., Reingold, E. M., & Schneider, B. A. (2011). Effects of aging and noise on real-time spoken word recognition: Evidence from eye movements. *Journal of Speech, Language, and Hearing Research*, *54*(1), 243–262. https://doi.org/10.1044/1092-4388(2010/09-0233)

Bianchi, F., Santurette, S., Wendt, D., & Dau, T. (2016). Pitch Discrimination in Musicians and Non-Musicians: Effects of Harmonic Resolvability and Processing Effort. *JARO - Journal of the Association for Research in Otolaryngology*, *17*(1), 69–79. https://doi.org/10.1007/s10162-015-0548-2

Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *Quarterly Journal of Experimental Psychology*, *73*(9), 1431–1443. https://doi.org/10.1177/1747021820916726

Clopper, C. G., Pisoni, D. B., & Tierney, A. T. (2006). Effects of Open-Set and Closed-Set Task Demands on Spoken Word Recognition. *Journal of the American Academy of Audiology*, *17*(5), 331–349.

Emily Shannon Fu Foundation. (2012). *AngelSim* (1.07.01). Emily Shannon Fu Foundation.

Farris-Trimble, A., & McMurray, B. (2013). Test–Retest Reliability of Eye Tracking in the Visual World Paradigm for the Study of Real-Time Spoken Word Recognition. *Journal of Speech Language and Hearing Research*, *56*(4), 1328. https://doi.org/10.1044/1092-4388(2012/12-0145)

Farris-Trimble, A., McMurray, B., Cigrand, N., & Tomblin, J. B. (2014). The process of spoken word recognition in the face of signal degradation. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(1), 308–327. https://doi.org/10.1021/nn300902w.Release

Feuerstein, J. F. (1992). Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort. *Ear and Hearing*, *13*(2), 80–86. https://doi.org/10.1097/00003446-199204000-00003

Francis, A. L., & Love, J. (2019). Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdisciplinary Reviews: Cognitive Science*, *February 2019*, 1–27. https://doi.org/10.1002/wcs.1514

Gagné, J. P., Besser, J., & Lemke, U. (2017). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, *21*, 1–25. https://doi.org/10.1177/2331216516687287

Hannagan, T., Magnuson, J. S., & Grainger, J. (2013). Spoken word recognition without a TRACE. *Frontiers in Psychology*, *4*(SEP), 1–17. https://doi.org/10.3389/fpsyg.2013.00563

Herrmann, B., & Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hearing Research*, *397*, 108016. https://doi.org/10.1016/j.heares.2020.108016

Hornsby, B. W. Y. (2013). The Effects of Hearing Aid Use on Listening Effort and Mental Fatigue Associated With Sustained Speech Processing Demands. *Ear and Hearing*, *34*(5), 523–534. https://doi.org/10.1097/AUD.0b013e31828003d8

Just, M. A., & Carpenter, P. A. (1992). A Capacity Theory of Comprehension: Individual Dirrefences in Working Memory. *Psychological Review*, *99*(1), 122–149.

Karatekin, C., Couperus, J. W., & Marcus, D. J. (2004). Attention allocation in the dual-task paradigm as measured through behavioral and psychophysiological responses. *Psychophysiology*, *41*(2), 175–185. https://doi.org/10.1111/j.1469-8986.2004.00147.x

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

Koelewijn, T., Shinn-Cunningham, B. G., Zekveld, A. A., & Kramer, S. E. (2014). The pupil response is sensitive to divided attention during speech processing. *Hearing Research*, *312*, 114–120. https://doi.org/10.1016/j.heares.2014.03.010

Koelewijn, T., Versfeld, N. J., & Kramer, S. E. (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. A., Lunner, T., & Kramer, S. E. (2018). The effect of reward on listening effort as reflected by the pupil dilation response. *Hearing Research*, *367*, 106–112. https://doi.org/10.1016/j.heares.2018.07.011

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in

Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lau, M. K., Hicks, C., Kroll, T., & Zupancic, S. (2019). Effect of auditory task type on physiological and subjective measures of listening effort in individuals with normal hearing. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1549–1560. https://doi.org/10.1044/2018_JSLHR-H-17-0473

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience*, *17*(3), 347–356. https://doi.org/10.1038/nn.3655

Mackersie, C. L., & 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

MATLAB 2017a. (2017). The Mathworks, Inc.

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin and Review*, *21*(3), 748–754. https://doi.org/10.3758/s13423-013-0544-7

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*(2), 145–160. https://doi.org/10.1016/j.jml.2011.04.004

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2017). Pupillometry reveals changes in physiological arousal during a sustained listening task.

Psychophysiology, *54*(2), 193–203. https://doi.org/10.1111/psyp.12772

McMurray, B. (2019). *EyelinkAnal* (4.11). https://osf.io/c35tg

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

McMurray, B., & Jongman, A. (2012). What information is necessary for speech categorization... *Psychological Review*, *118*(2), 219–246. https://doi.org/10.1037/a0022325.What

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*, B33–B42.

Mitterer, H., & Mattys, S. L. (2017). How does cognitive load influence speech perception? An encoding hypothesis. *Attention, Perception, & Psychophysics*, *79*(1), 344–351. https://doi.org/10.3758/s13414-016-1195-3

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian Model of Continuous Speech Recognition. *Psychological Review*, *115*(2), 357–395. https://doi.org/10.1037/0033-295X.115.2.357

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective and Behavioral Neuroscience*, *5*(3), 263–281. https://doi.org/10.3758/CABN.5.3.263

Ohlenforst, B., Zekveld, A. A., Lunner, T., Wendt, D., Naylor, G., Wang, Y., Versfeld, N. J., & Kramer, S. E. (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. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear & Hearing*, *39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & 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

Picou, E. M., & Ricketts, T. A. (2014). The effect of changing the secondary task in dual-task paradigms for measuring listening effort. *Ear and Hearing*, *35*(6), 611–622. https://doi.org/10.1097/AUD.0000000000000055

Picou, E. M., & Ricketts, T. A. (2018). The relationship between speech recognition, behavioural listening effort, and subjective ratings. *International Journal of Audiology*, *57*(6), 457–467. https://doi.org/10.1080/14992027.2018.1431696

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2011). Visual Cues and Listening Effort: Individual Variability. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1416–1430. https://doi.org/10.1044/1092-4388(2011/10-0154)

Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D. W., Slotow, R., & Hamer, M. (2014). Using commonality analysis in multiple regressions: A tool to decompose regression effects in the face of multicollinearity. *Methods in Ecology and Evolution*, *5*(4), 320–328. https://doi.org/10.1111/2041-210X.12166

Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective Measures of Listening

Effort: Effects of Background Noise and Noise Reduction. *Journal of Speech Language and Hearing Research*, *52*(5), 1230. https://doi.org/10.1044/1092-4388(2009/08-0111)

Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal of Psychophysiology*, *52*(1), 77–86. https://doi.org/10.1016/j.ijpsycho.2003.12.005

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures. *Journal of Speech Language and Hearing Research*, *61*(6), 1463. https://doi.org/10.1044/2018_JSLHR-H-17-0257

Strand, J. F., Ray, L., Dillman-Hasso, N. H., Villanueva, J., & Brown, V. A. (2021). Understanding Speech amid the Jingle and Jangle: Recommendations for Improving Measurement Practices in Listening Effort Research. *Auditory Perception & Cognition*, 1–20. https://doi.org/10.1080/25742442.2021.1903293

Strauss, D. J., & Francis, A. L. (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

Thompson, B. (1995). Stepwise regression and stepwise discriminant analysis need not apply here. *Educational and Psychological Measurement*, *55*(4), 525–534.

Ullsperger, M., Harsay, H. A., Wessel, J. R., & Ridderinkhof, K. R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Structure & Function*, *214*(5–6), 629–643. https://doi.org/10.1007/s00429-010-0261-1

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin and Review*, 1–11.

https://doi.org/10.3758/s13423-018-1432-y

Wagner, A. E., Toffanin, P., & Baskent, D. (2016). The timing and effort of lexical access in natural and degraded speech. *Frontiers in Psychology*, *7*(MAR), 1–14. https://doi.org/10.3389/fpsyg.2016.00398

Ward, K. M., Shen, J., Souza, P. E., & Grieco-Calub, T. M. (2017). Age-Related Differences in Listening Effort during Degraded Speech Recognition. *Ear and Hearing*, *38*(1), 74–84. https://doi.org/10.1097/AUD.0000000000000355

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology*, *7*(MAR), 1–12. https://doi.org/10.3389/fpsyg.2016.00345

Whittingham, M. J., Stephens, P. A., Bradbury, R. B., & Freckleton, R. P. (2006). Why do we still use stepwise modelling in ecology and behaviour? *Journal of Animal Ecology*, *75*(5), 1182–1189. https://doi.org/10.1111/j.1365-2656.2006.01141.x

Winn, M. B. (2016). Rapid Release from Listening Effort Resulting from Semantic Context, and Effects of Spectral Degradation and Cochlear Implants. *Trends in Hearing*, *20*, 1–17. https://doi.org/10.1177/2331216516669723

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The Impact of Auditory Spectral Resolution on Listening Effort Revealed by Pupil Dilation. *Ear and Hearing*, *36*(4), 153–165. https://doi.org/10.1097/AUD.0000000000000145.The

Winn, M. B., & Moore, A. N. (2018). Pupillometry Reveals That Context Benefit in Speech Perception Can Be Disrupted by Later-Occurring Sounds, Especially in Listeners With Cochlear Implants. *Trends in Hearing*, *22*, 1–22. https://doi.org/10.1177/2331216518808962

Winn, M. B., & Teece, K. H. (2020). *Listening Effort is not the same as Speech Intelligibility Score*. https://doi.org/10.31234/osf.io/vk65w

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in Hearing*, *22*, 233121651880086. https://doi.org/10.1177/2331216518800869

Wu, Y.-H., Stangl, E., Zhang, X., Perkins, J., & Eilers, E. (2016). Psychometric Functions of Dual-Task Paradigms for Measuring Listening Effort. *Ear and Hearing*, *37*(6), 660–670. https://doi.org/10.1097/AUD.0000000000000335

Zekveld, A. A., Heslenfeld, D. J., Johnsrude, I. S., Versfeld, N. J., & Kramer, S. E. (2014). The eye as a window to the listening brain: Neural correlates of pupil size as a measure of cognitive listening load. *NeuroImage*, *101*, 76–86. https://doi.org/10.1016/j.neuroimage.2014.06.069

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: Insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. https://doi.org/10.1111/psyp.12151

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

Zekveld, A. A., van Scheepen, J. A. M., Versfeld, N. J., Veerman, E. C. I., & Kramer, S. E. (2019). Please try harder! The influence of hearing status and evaluative feedback during listening on the pupil dilation response, saliva-cortisol and saliva alpha-amylase levels. *Hearing Research*, *381*, 107768. https://doi.org/10.1016/j.heares.2019.07.005