Impaired Performance in Noise: Disentangling Listening Effort from the Irrelevant Speech Effect

Janna Wennberg[1], Naseem Dillman-Hasso[2], Violet A. Brown[3], Julia F. Strand[4]

1. University of California, San Diego, Department of Psychology
2. The Ohio State University, School of Environment and Natural Resources
3. Washington University in St. Louis, Department of Psychological & Brain Sciences
4. Carleton College, Department of Psychology

Acknowledgements: This work was supported by the National Science Foundation through a Graduate Research Fellowship awarded to Violet Brown (DGE-1745038), a National Institutes of Health grant via the National Institute on Deafness and Communication Disorders awarded to Julia Strand (R15-DC018114), and Carleton College. The authors have no conflicts of interest. Correspondence regarding this paper should be addressed to jwennber@ucsd.edu or to jstrand@carleton.edu.

Abstract

Noise can reduce the intelligibility of spoken language and increase the effort necessary to understand speech. *Listening effort,* “the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task” (Pichora-Fuller et al., 2016), is commonly assessed by measuring response times to secondary tasks while listening to speech or by testing memory for the content of the speech. Increasing the level of background noise tends to slow responses and impair memory, and these effects are attributed to the resource-intensive process of reevaluating speech that was initially obscured or misheard. However, given that noise can impair performance on cognitive tasks that do not require processing auditory information, it is possible that noise-induced impairments typically ascribed to processing degraded speech may instead reflect increased cognitive load from the presence of noise itself. The current study will assess whether noise, in the absence of a speech task, can affect performance on tasks intended to measure listening effort. In Experiment 1 (positive control), target speech consisting of single words will be presented aurally in background noise and we will measure listening effort with three commonly-used paradigms. Experiment 2 is identical except that the target words will be presented orthographically rather than aurally. If tasks intended to measure listening effort are affected by the presence of background noise even in the absence of auditory speech, it would suggest that some effects typically ascribed to listening effort from processing degraded speech may be attributable to other forms of cognitive interference. Including multiple measures of listening effort will help to clarify which measures of effort seem to be particularly affected by noise-induced cognitive interference.

Keywords: listening effort, speech perception, irrelevant sound effect

Impaired Performance in Noise: Disentangling Listening Effort from the Irrelevant Speech Effect

Anyone who has attended a noisy sporting event or cocktail party is familiar with the challenges of listening to speech in noise. Not only does background noise lead to poorer performance on transcription or identification tasks (Pisoni, 1996), but it can also lead to increases in *listening effort*: “the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task” (Pichora-Fuller et al., 2016). Although the definition of listening effort outlined above applies to listening tasks broadly, the majority of research on listening effort has focused on the cognitive resources necessary to process speech. In the speech literature, listening effort is commonly measured using behavioral tasks in which participants listen to speech and either simultaneously perform another task (dual-task paradigms; see Gagné et al., 2017 for a review) or store the speech in memory for later recall (recall paradigms; e.g., McCoy et al., 2005; Rabbitt, 1968). These measures rely on the assumption that humans have a finite pool of cognitive resources (Kahneman, 1973), so when more resources are needed to process the speech, fewer remain available to efficiently perform other tasks.

According to the Ease of Language Understanding (ELU; Rönnberg et al., 2013) model, any situation that elicits a mismatch between the incoming acoustic signal and representations of words in the listener’s memory (e.g., background noise) will increase listening effort because additional cognitive resources such as working memory must be recruited to resolve these mismatches. Critically, the detrimental effects of noise or other types of signal degradation on listening effort are assumed to stem from the challenges of processing the speech in difficult listening conditions—that is, from the resource-intensive process of reevaluating phonemes or words that were initially obscured or misheard. A common method for degrading speech to experimentally induce listening effort is to add or increase the level of masking noise. Noise has been shown to affect multiple tasks intended to measure listening effort; it slows response times in dual-task paradigms (e.g., Picou & Ricketts, 2014), impairs performance on recall tasks (e.g., see Brown & Strand, 2019a; Picou & Ricketts, 2014; Rabbitt, 1968), affects pupillary responses (see Van Engen & McLaughlin, 2018), and increases subjective ratings of listening effort (e.g., Johnson et al., 2015; Strand et al., 2018).

Noise-induced changes in tasks intended to measure listening effort are typically thought to reflect the cognitive challenges of parsing degraded speech. However, there is another possible explanation: these effects may reflect increased cognitive load from the presence of noise itself, rather than from the effect of noise on processing spoken words. Indeed, even cognitive tasks that do not require processing acoustic speech can be negatively affected by the presence of noise (see Szalma & Hancock, 2011). For example, background noise impairs performance on the Raven's Progressive Matrices test (Dobbs et al., 2011), affects verbal reasoning (Dobbs et al., 2011), and hinders lipreading performance (Campbell et al., 2002; Myerson et al., 2016).

One area in which the negative effects of noise on cognitive performance are particularly well-documented is in the memory literature: Noise impairs performance on memory tasks even when the noise is not relevant to the task (e.g., the task involves visual memory; Wais & Gazzaley, 2011; Weisz & Schlittmeier, 2006). This phenomenon—known as the irrelevant speech effect[1] (Baddeley & Salamé, 1986; Beaman et al., 1998; Beaman & Jones, 1997; Norris et al., 2004; Salamé & Baddeley, 1982)—may occur because noise (and speech in particular) is automatically processed in the phonological loop, which disrupts working memory for other information that is being rehearsed subvocally (e.g., Neath, 2000; Salamé & Baddeley, 1982). Given that verbal stimuli from both auditory and visual modalities are processed in the phonological loop, aurally presented speech stimuli can impair recall of to-be-remembered items, even when those items are presented visually. For example, Colle and Welsh (1976) demonstrated that recall of visually presented digits was impaired when an unfamiliar language was played in the background. Critically, noise-induced interference in memory tasks does not occur for

[1] Note that the term *irrelevant speech effect* has also been broadened to include non-speech sounds (i.e., the *irrelevant sound effect*), reflecting the fact that modulating non-speech sounds such as changing tones can also interfere with working memory (Jones et al., 1999; Jones & Macken, 1993).

all types of noise; physically changing sound such as speech and music—but not steady-state noise—disrupts serial recall of visually-presented items (this is known as the changing state hypothesis; Jones & Macken, 1993; Jones & Morris, 1992). In sum, there is ample evidence that modulating noise can impair performance on memory tasks, even cross-modally.

To be clear, noise-induced cognitive interference cannot explain the entirety of the listening effort literature. First, many factors other than background noise have been shown to affect listening effort. For example, vocoded speech (Winn, 2016) and reverberant speech (Rennies et al., 2014) lead to greater listening effort than natural speech, and nonnative-accented speech requires greater listening effort than native-accented speech (Borghini & Hazan, 2018; Brown et al., 2020; McLaughlin & Van Engen, 2020). Second, steady-state noise increases listening effort (e.g., Brown & Strand, 2018) without producing interference on other cognitive tasks, demonstrating that not all findings in the listening effort literature can be attributed to the irrelevant speech effect. Nevertheless, adding background noise is a common way to induce listening effort, and the mechanism underlying how noise impairs performance on tasks intended to measure listening effort remains unclear. Slower response times and poorer recall for speech in noise may indeed reflect cognitive challenges associated with resolving mismatches between input and mental representations (as is typically assumed in the listening effort literature). Alternatively or in addition, these effects may be driven by the fact that noise impairs performance on cognitive tasks more generally. The existing work is not able to fully distinguish between these possibilities.

A recent study attempted to assess whether performance on a task commonly assumed to measure listening effort was affected by the presence of background noise in the absence of speech (Brown & Strand, 2018). In this task, participants made speeded judgments about visually-presented numbers while also listening to and repeating aurally-presented words (Picou & Ricketts, 2014; Sarampalis et al., 2009). In line with prior work, Brown and Strand (2018) demonstrated slower response times to the number judgment task as noise level increased. However, when the same task was performed without the aurally-presented speech, the noise-induced slowdowns on the number judgment task were not observed. This suggests that when the background noise was loud and speech was present, the slowed responses were a function of increased effort from processing degraded speech; the authors did not observe cognitive interference from the presence of noise itself.

However, the finding that noise alone does not impair performance on tasks intended to measure listening effort may not extend to other listening effort paradigms. Listening effort has been assessed using a variety of tasks, and there is growing doubt that those measures tap into the same underlying construct (Alhanbali et al., 2019; Strand et al., 2018). Indeed, measures of listening effort are often weakly intercorrelated with one another (e.g., Johnson et al., 2015; Seeman & Sims, 2015; Strand et al., 2018), and they produce different patterns of results even when the same noise conditions and speech stimuli are used (Brown & Strand, 2019a). It is therefore possible that some tasks designed to measure listening effort are more affected by the presence of background noise than the dual-task paradigm used by Brown and Strand (2018). Specifically, given that irrelevant sounds are particularly detrimental to memory tasks (e.g., Jones & Macken, 1993; Salamé & Baddeley, 1982), measures of listening effort that rely on encoding and recalling information may be more susceptible to noise-induced cognitive interference than dual-task paradigms.

Further, the results of Brown and Strand (2018) may not generalize to other types of noise. That study was conducted in steady-state background noise, which is commonly used to degrade speech by masking phonetic detail. However, steady-state noise is not as disruptive to cognitive, perceptual, and motor tasks as complex forms of noise (see Szalma & Hancock, 2011 for a meta-analysis), consistent with the changing-state hypothesis (Jones & Macken, 1993; Jones & Morris, 1992). For example, noise with temporal variation is especially detrimental to working memory performance (Jones et al., 1990; Salamé & Baddeley, 1989; Tremblay et al., 2001), and two-talker babble interferes with lipreading ability more than steady-state noise does (Lidestam et al., 2014; Myerson et al., 2016). Thus, although Brown and

Strand (2018) provided evidence that steady-state noise in the absence of speech did not impair performance on a particular listening effort task, it is not clear whether that finding would extend to other tasks or other types of masking noise.

The Current Study

The goal of the current study was to distinguish between two explanations for noise-induced performance decrements on tasks intended to measure listening effort. One explanation is that noise leads to mismatches between the incoming speech and representations of words stored in memory, and cognitive resources (i.e., listening effort) must be recruited to resolve these mismatches, consistent with the ELU account (Rönnberg et al., 2013). Another explanation is that noise may increase cognitive load and therefore interfere with working memory, consistent with research on the irrelevant speech effect (Baddeley & Salamé, 1986). Either mechanism alone, or both explanations jointly, could explain the finding that performance on listening effort tasks is impaired by noise. The current research aims to disentangle these explanations and provide insights about the mechanism by which noise affects performance on tasks intended to measure listening effort.

The present study employed three widely-used listening effort tasks in which participants listen to and repeat words in silence, steady-state speech-shaped noise, and two-talker babble. Noise was always presented aurally, and the target stimuli were presented either aurally (Experiment 1) as they are in traditional listening effort tasks, or orthographically (Experiment 2) such that the tasks closely mirror listening effort tasks, but noise effets cannot be attributable to mismatches between acoustic input and representations in memory. Experiment 1 serves as a positive control to ensure that we can demonstrate effects typically attributable to listening effort using each of the three paradigms, and Experiment 2 will enable us to assess the extent to which performance impairments on the three tasks are due to noise-related cognitive interference generally rather than listening effort resulting from processing degraded speech.

The three tasks we selected are well established in the listening effort literature: a vibrotactile dual-task paradigm that requires making judgments about the duration of vibrations presented to the index finger while listening to and repeating words (Brown & Strand, 2019a; Fraser et al., 2010; Gosselin & Gagné, 2011), a verbal dual-task paradigm that involves judging whether isolated words are nouns (Picou & Ricketts, 2014; Strand et al., 2018), and a running memory task that requires storing and later recalling lists of words (McCoy et al., 2005; Sommers & Phelps, 2016; Strand et al., 2018). We selected these tasks for three reasons. First, we wanted to include both a dual-task and a recall paradigm, as they are the most commonly used behavioral methods of assessing listening effort. Second, we wanted to include a dual-task paradigm that requires verbal processing (making a judgement about a word) to enable a more direct comparison to the recall task, which also requires verbal processing. Had we only included the non-verbal dual-task paradigm (the vibrotactile task) and found that background noise affected performance on the recall and dual-task paradigms differently, it would be unclear whether the observed differences were the result of a difference between recall and dual-tasks or between verbal and non-verbal tasks. Third, given that successfully performing the noun judgment task relies on accurately perceiving the speech (i.e., a participant cannot judge a word as a noun if they did not hear it), we wanted to include a dual-task paradigm in which the speech task and listening effort task can be completed independently. Each task was completed in silence, steady-state noise (to induce masking at the level of the auditory periphery—energetic masking), and two-talker babble (to induce masking attributable to higher-level cognitive processing—informational masking; Freyman et al., 1999) (see Figure 1).

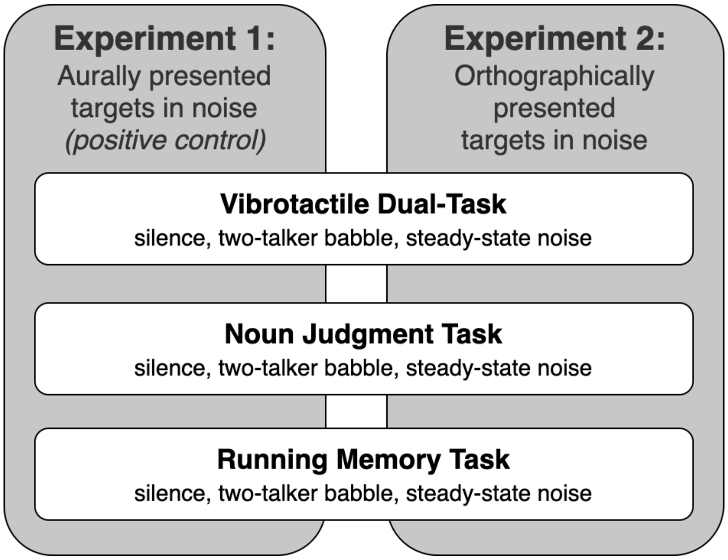
Experiment 1:
Aurally presented targets in noise
(positive control)
Experiment 2:
Orthographically presented targets in noise
Vibrotactile Dual-Task
silence, two-talker babble, steady-state noise
Noun Judgment Task
silence, two-talker babble, steady-state noise
Running Memory Task
silence, two-talker babble, steady-state noise

Figure 1. Schematic of experimental procedure. Participants completed all three tasks and all three noise conditions in either Experiment 1 or Experiment 2.

Finally, to obtain a measure of self-reported listening difficulty, participants completed the four questions on the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988) intended to measure mental demand, performance, effort and frustration. We excluded the questions regarding physical and temporal demand because they are not relevant to our research question and may confuse participants. These ratings were obtained throughout each of the three tasks in both experiments.

Hypotheses

Experiment 1

In Experiment 1, the target stimuli were presented aurally, as is always the case with listening effort research. Experiment 1 therefore serves as a positive control to ensure that we can demonstrate results consistent with prior work on listening effort using these particular tasks and types of noise.

Hypothesis 1: *We expected to replicate the well-established finding that both steady-state noise and two-talker babble adversely affect performance on all three tasks relative to listening in quiet.* We did not have specific predictions about whether two-talker babble or steady-state noise would lead to greater increases in listening effort because the extent of effort depends on idiosyncrasies of the stimuli such as characteristics of the talkers and signal-to-noise ratio. The magnitude of the effects from Experiment 1 will serve as a point of comparison for Experiment 2.

Experiment 2

The purpose of Experiment 2 was to assess whether the performance deficits observed in Experiment 1 might be driven by noise-induced cognitive interference rather than listening effort associated with processing degraded speech. Experiment 2 was identical to Experiment 1 except that the target words were presented orthographically rather than aurally. If background noise impairs performance on the tasks in Experiment 2, it would suggest that findings typically attributed to difficulty with listening to speech may stem from cognitive challenges associated with noise. Our hypotheses for Experiment 2 were as follows:

Hypothesis 2a: *We expected that steady-state noise would not impair performance relative to performance on the same tasks in silence.* Brown and Strand (2018) found that response times to a non-verbal dual-task paradigm were unaffected by the level of steady-state noise, so we expected that performance on the vibrotactile task (another non-verbal dual-task paradigm) is unlikely to be affected by steady-state noise. Furthermore, the changing state hypothesis predicts that modulating noise (such as speech) taxes the phonological loop, but steady-state noise does not (Jones et al., 1990; Salamé & Baddeley, 1987; Weisz & Schlittmeier, 2006). We therefore expected that performance on the noun judgment and recall tasks would also be unaffected by steady-state noise. If we find that steady-state noise impairs performance on the recall task, this would be inconsistent with the changing state hypothesis (Jones & Macken, 1993; Jones & Morris, 1992) but would be consistent with the general finding that irrelevant sounds can impair performance on cognitive tasks. Although we do not anticipate this effect, this would indicate that tasks intended to measure listening effort may also be affected by cognitive challenges associated with the presence of noise.

Hypothesis 2b: *We expected that performance on all three tasks would be negatively affected by the presence of two-talker babble relative to performance in silence.* Baddeley's model of working memory would predict that two-talker babble automatically enters the phonological loop and interferes with processing and remembering words (see Baddeley, 1992). This leaves fewer resources available for completing the other tasks, leading to slower response times and poorer recall of orthographic words when the task is completed in two-talker babble relative to silence. Given research on the irrelevant speech effect, finding no difference between the two-talker babble and silence conditions for any task (and the recall task in particular) would be surprising and would indicate that the two-talker babble did not place sufficient strain on the phonological loop to impair performance. This would strengthen the argument that these listening effort tasks are assessing the effort that results from processing degraded speech rather than from the mere presence of background noise.

Hypothesis 2c: *We expected that the magnitude of the interference from the two-talker babble would be largest for the recall task, smaller for the noun judgment task, and smallest for the vibrotactile dual-task.* We anticipated that two-talker babble would be particularly taxing for tasks that require greater recruitment of working memory (the recall task) and those that require more verbal processing (the noun judgment task). This will clarify which tasks assess the listening effort associated with processing degraded speech and which tasks assess more general noise-induced cognitive interference.

Experiment 1: Auditory Presentation (Positive Control)

All stimuli, raw data, and code for statistical analyses are available at https://osf.io/asnqj/.

Method

Participants

Participants will consist of 55 young adults (ages 18-28) from the Carleton College community with self-reported normal or corrected-to-normal vision and no known hearing impairment. We will recruit participants via posted advertisements, word of mouth, and email. Sample size was pre-determined via power analysis (see details below). All procedures will be approved by the Carleton College Institutional Review Board, and participants will give written consent prior to participating. Participants in both experiments will receive $17 for 90 minutes of participation.

Speech Stimuli

Speech stimuli consisted of 810 words selected from the English Lexicon Project (Balota et al., 2007), and each word's dominant part of speech was determined from the SUBTLEX-US database (Brysbaert et al., 2012). Following the conventions of Strand, Brown, and Barbour (2020), we subsetted the full database to only include words with log-frequencies of three or higher (Brysbaert & New, 2009), two to five phonemes, and one or two syllables. We also excluded proper nouns, articles, conjunctions, interjections, profane or emotionally evocative words, and homographs (e.g. "lead"). We then selected words at random from this set, replacing multiple instances of homophones and words with multiple forms (e.g. "bad" and "badly"). Finally, we replaced words as needed until 55% of words were classified predominantly as nouns to follow the norms of previous studies that have used the noun judgment task (Picou & Ricketts, 2014; Strand et al., 2018). Of the 810 words, we randomly assigned 180 to appear in the vibrotactile task, 270 to appear in the noun judgment task, and 360 to appear in the recall task, maintaining the 55% noun composition in each of the three tasks. The number of stimuli presented in each task was determined to ensure that the analyses associated with each task were sufficiently powered (see Power Analysis below). The three sets of words were matched on length, frequency, number of orthographic and phonological neighbors, number of syllables, and number of phonemes.

Speech stimuli will be recorded with a Blue Yeti microphone by a female speaker without a strong regional accent and will be edited and equated on root-mean-square amplitude using Adobe Audition. Speech will be presented in noise at a signal-to-noise ratio of approximately -6 dB, but the signal-to-noise ratio will be set independently for two-talker babble and steady-state noise. We will conduct pilot testing to determine a level of noise that resulted in intelligibility levels of approximately 70% correct in each noise condition, and use those levels for all participants. This level of difficulty will be chosen to make the task difficult enough for noise effects to emerge but not so difficult that participants can not hear the speech and therefore can not complete the tasks. The signal-to-noise ratio will be set separately for both noise conditions to ensure that they result in similar levels of word recognition performance (note that given amplitude fluctuations in the two-talker babble, the momentary SNR for this noise condition will vary). Noise files will be played at approximately 68 dB SPL throughout the experiment, and speech levels will be set to attain the desired level of performance.

Noise Stimuli

The steady-state noise will consist of speech-shaped noise generated in Praat (version 6.0.36) to match the long-term average spectrum of the target stimuli (Winn, 2018). The two-talker babble will consist of two continuous speech streams in which each stream will be produced by a different female speaker. The two-talker babble will also match the long-term average spectrum of the target stimuli (see Brouwer et al., 2012). Two-talker babble will consist of two female speakers producing the BKB (Bamford-Kowal-Bench; Bench et al., 1979) sentences, which are simple, meaningful sentences (e.g., “The clown had a funny face”). RMS amplitude equalized recordings of these BKB sentences were obtained from Van Engen (2010). To generate the two-talker babble, we will combine the audio files for each speaker into one continuous stream, then overlay the two speakers streams. Natural fluctuations in speaking speed ensure that sentences do not consistently start and stop at the same time, and the original stimuli are created such that there are no pauses between sentences.

For all tasks, noise will be played continuously during blocks with background noise present, but will pause when participants are completing the NASA-TLX. We chose continuous noise presentation for the vibrotactile and noun judgement tasks in order to increase temporal uncertainty, thus making the listening task more challenging. Although running memory tasks typically do not have noise during the recall portion, we chose to include it here because continuous noise presentation for some tasks but not for others could result in different degrees of noise habituation across tasks, further limiting our ability to compare effects of noise across these tasks.

NASA-TLX

Participants completed four of the six questions from the NASA-TLX throughout the three tasks. Participants responded by clicking a location along an unnumbered 21-point scale ranging from "Very Low" (or "Failure" in the case of the performance question) to "Very High" (or "Perfect" in the case of the performance question).[2] Each question was presented in the following order:

1. "How mentally demanding was the task?" [mental demand]
2. "How successful were you in accomplishing what you were asked to do?" [performance]
3. "How hard did you have to work to accomplish your level of performance?" [effort]
4. "How insecure, discouraged, irritated, stressed, and annoyed were you?" [frustration]

We will only analyze data for the effort question, but in order to isolate subjective effort from other factors (such as beliefs about performance), we included the other questions.

Procedure

Participants will sit a comfortable distance from a 21.5-inch iMac computer and will wear noise-cancelling headphones to attenuate sounds from the vibrotactile apparatus as well as sounds outside the testing environment. All participants will complete three tasks: a vibrotactile pattern recognition task, a noun judgment task, and a running memory task (see Figure 1). Each task will be conducted in three

[2] Note that the original survey presented "Perfect" on the left end of the performance scale and "Failure" on the right end, but we switched the scale limits to be consistent with the other questions on the NASA-TLX.

noise conditions: silence, steady-state noise, and two-talker babble. The three tasks will be blocked and the order of the blocks will be randomized. Within each task, words will be randomly divided into three lists to be presented in each of the noise conditions to ensure that within a task, words appear in all conditions approximately the same number of times (albeit for different participants). Within each task, the noise conditions will be blocked, and the order in which they are presented will also be randomized. For all tasks, participants will repeat words aloud as they hear them, and responses will be recorded in Audacity and coded for identification accuracy offline by research assistants.

Each participant will complete the NASA-TLX four times per noise condition to provide multiple observations per participant per condition. Thus, the NASA-TLX will be completed every 15 words for the vibrotactile task, every 22 words for the noun judgment task (with the exception of the last presentation, which will happen after 24 words), and every 30 words for the recall task.

Vibrotactile task. The vibrotactile task will be identical to the one used in two previous studies by our lab (Brown & Strand, 2019a, 2019b). Vibrotactile stimulation will be presented via a custom-made apparatus consisting of a 3D-printed finger rest and a direct current vibrating motor that delivers pulse trains of various lengths. Vibrations will be delivered to the index finger of the participant's non-dominant hand. During the task, participants will be presented with short (100 ms), medium (150 ms), and long (250 ms) pulses from the vibrotactile device. The apparatus and the participant's hand will be placed inside a box lined with noise-attenuating foam to reduce any sounds generated from the vibrating apparatus. Prior to the main task, all participants will complete a familiarization block. Participants will first be presented with two short pulses, two medium pulses, and two long pulses. During familiarization, participants will identify 18 randomly ordered pulses (six of each length) by pressing the appropriate button on the box. In the event of an incorrect response, the correct answer will immediately be displayed on the screen. In order to pass the familiarization phase, participants must obtain at least 75% accuracy (14 out of the 18 pulses). If this threshold is not met, the entire familiarization block will be repeated.

Following successful completion of the familiarization block, participants will complete the remaining three experimental blocks.

Participants will complete a total of 180 trials in the main task (60 per noise condition). Each trial will consist of a vibrotactile pulse and an auditorily presented word. The pulse will begin somewhere between 100 ms before the onset of the word and 150 ms after the onset of the word, randomly selected from 50 ms intervals. We chose these onset times because they ensure that the cognitive processing required to perform the speech task and the vibrotactile pulse classification task coincide. That is, even if the shortest pulse is presented at the earliest onset time and the pulse itself does not coincide with presentation of the word, making a judgment about the pulse will coincide with presentation of the word. After the word was presented, there will be a variable interstimulus interval ranging from 2,500 to 3,500 ms, randomly selected from 250 ms intervals. Participants will respond to the vibrotactile stimulus by pressing the appropriate button on a button box and will then repeat the word aloud. The outcome measure of interest will be response time to make the vibration length judgment, measured from the onset of the vibration.

Noun judgment task. The noun judgment task will follow the conventions of previous work implementing this task (Picou & Ricketts, 2014; Strand et al., 2018). At the start of each trial, a word will be presented through headphones, and participants will be asked to press a button on a button box as quickly as possible if the word can ever be classified as a noun. After making this noun judgment, participants will repeat the word they perceived aloud, regardless of whether they had classified it as a noun. The interstimulus interval will again range from 2,500 to 3,500 ms in random 250 ms intervals. The outcome measure of interest will be response time on trials during which the participant indicated that the word was a noun. Following the conventions of prior work, we will analyze all noun responses as opposed to all "correct" responses because many nouns can be categorized as other parts of speech (see

Picou & Ricketts, 2014; Strand et al., 2018). Participants will complete a total of 270 trials in the main task (90 per noise condition).

Running memory task. The procedures for this task will be similar to those we employed in our previous work (Brown & Strand, 2019a). Participants will complete three blocks of the running memory task (McCoy et al., 2005; Morris & Jones, 1990; Sommers et al., 2015; Sommers & Phelps, 2016; Strand et al., 2018), one per noise condition. In each block, participants will be presented with 16 lists of words ranging in length from five to ten words, with 1,000 ms between each word. Each list length will be presented three times per noise condition, with the exception of the shortest and longest list lengths (5 and 10 words), which will each be presented twice, for a total of 120 words per condition. Words were assigned to lists randomly, but each list was manually checked to ensure that none of the words within a list were semantically related. Participants will be instructed to repeat each word aloud immediately after presentation, and at the end of each list, verbally recall the last four words in each list in any order. The next trial will be initiated after a button press or after eight seconds have elapsed. We will include all words when calculating intelligibility scores for exclusion purposes (see below), but the outcome measure of interest will be recall of the words in the 3- and 4- back positions of each list. With 16 lists of words per noise condition, this will result in 32 critical items per noise condition. Words will be counted as having been recalled correctly if they are recalled as they were perceived—that is, if the participant initially misperceives a word but then later recalls that misperception, this will be counted as correct (e.g., Brown & Strand, 2019a; Johnson et al., 2015; Pichora-Fuller et al., 1995).

Exclusion Criteria

Participants will be excluded from all analyses in the event of unresolvable equipment issues during the experiment (e.g., computer crashes) or if a participant reports doing the task incorrectly (e.g., taking off the headphones during the experiment).

We will also exclude participants based on task performance. For each task, we will determine each participant's accuracy at repeating the target words aloud to ensure that we only include participants who attend to the primary speech task. Particularly for the dual-task paradigms, if participants stop attending to the speech task, this may result in performance improvements on the secondary tasks. If, for any task and any noise condition, a participant's word identification accuracy is more than three standard deviations below the mean accuracy for that condition, that participant's data will be excluded from all analyses for that task. However, if a participant meets this exclusion criterion in a condition but still achieves at least 90% accuracy for that condition, they will not be excluded. We made this decision because if performance is at ceiling level in a given condition, this renders high means and small standard deviations, which can result in an extremely conservative performance cutoff. We wanted to avoid unnecessarily discarding data from participants who were performing the task reasonably well, so we will remove participants only if their performance is below 90% in addition to being three standard deviations below the mean. Finally, participants will be excluded from a response time task if their mean response time in any condition is more than three standard deviations above or below the mean for that condition. Note that if a participant meets an exclusion criterion in any of the three noise conditions for a particular task, their data will be excluded from all noise conditions, but only for that task.

Individual trials will be excluded if response times to the noun judgment or vibrotactile task are more than three median absolute deviations above or below that participant's median response time. Following the recommendations of Leys and colleagues (2013), we will use medians and median absolute deviations in lieu of means and standard deviations here because raw response times tend to be skewed.

Power Analysis

We will collect data from 55 participants for Experiment 1. Assuming conservative response rates of 0.75 for the vibrotactile task and 0.50 for the noun judgement task (in which a button is pressed only when a word can be a noun), we expect to obtain 45 trials per participant per noise condition for both the vibrotactile task (0.75 response rate * 60 total words) and the noun judgement task (0.50 response rate * 90 total words). Collecting data from 55 participants will give us 2,475 total response time trials for each experimental condition in both response time tasks.

To ensure that 55 participants would power us for the effects of interest, we performed a frequentist *a priori* power analysis using the *simr* package in R, using data from Brown and Strand (2019a) to establish meaningful model estimates. Participants in Brown and Strand (2019a) completed the same vibrotactile and running memory tasks that we propose in the present study and with very similar speech stimuli. Here, we assume that the power analysis using the vibrotactile response time data will generalize to response times from the noun judgement task proposed. Brown and Strand (2019a) used SNRs of -4 dB and 10 dB, whereas in the proposed study, we plan to use a more difficult SNR and silence. Thus, we expect that the effects observed will be as large as—or quite possibly larger than—those in Brown and Strand (2019a).

Participants in Brown and Strand (2019a) completed considerably more response time trials than participants in the proposed study. To avoid overestimating the power for the proposed study, we randomly sampled 45 trials per participant per noise conditions to include in our power simulations. We chose this number of trials because, in the proposed study, we will obtain an expected 45 trials per participant per noise condition. The running memory task in Brown and Strand (2019a) had the same number of critical trials per condition as proposed in the present study, so we did not downsample.

For response time and recall data separately, we obtained a full power curve by computing power at a range of effect sizes, rather than computing a single estimate (Morey et al., 2021). The power curves are shown in Figure 2. When building our models, we attempted to include by-word random slopes for noise, but this led to unresolvable convergence issues. However, we still included by-participant random slopes for noise, as well as by-participant and by-word random intercepts. Using estimates from this model, we ran 200 simulations at each of 10 effect sizes ranging from 2% to 20%, in increments of 2%.

Effect sizes are given as a percentage change: for the vibrotactile task, this is a percent increase in response time in louder noise relative to quieter noise (the effect size observed in Brown and Strand (2019a) was 10%). For the running memory task, this is a percent decrease in recall accuracy in louder noise relative to quieter noise (the effect size observed in Brown and Strand (2019a) was 15%). According to our simulations, we would be 100% powered to detect effect sizes of the same magnitude as those observed in Brown and Strand (2019a) with a sample size of 55 participants. Further, simulated power was well above 90% for effect sizes of roughly 7% or higher for the vibrotactile task and 9% or higher for the running memory task. Thus, collecting data from 55 participants will be sufficient even in the event that our effect sizes are smaller than those in Brown and Strand (2019a).

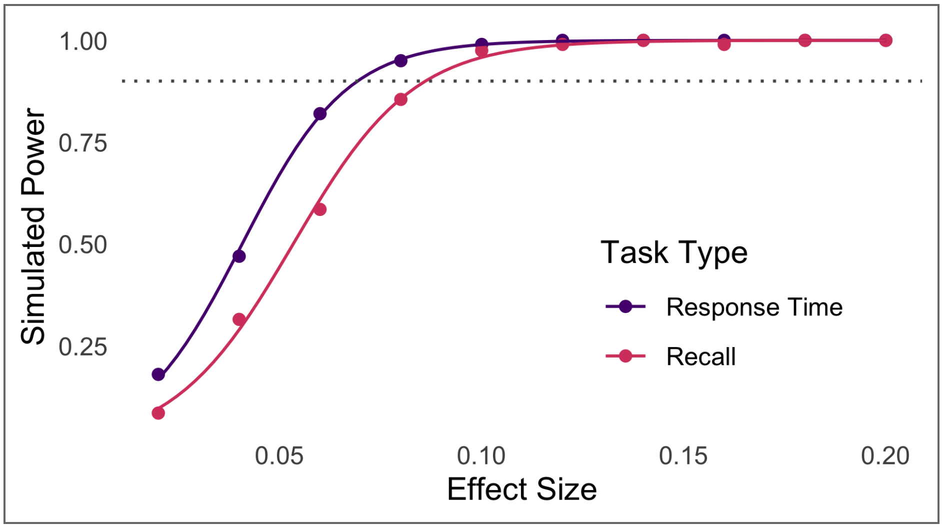

Figure 2. Power curve for Experiment 1. Curve was obtained from simulations with *simr* and data from Brown and Strand (2019), adjusted to 55 participants. The gray dotted line denotes power of 0.90.

Data Analysis

Model fitting. We will use both frequentist and Bayesian multilevel modeling to assess evidence for our hypotheses, and we will use the *lme4* (version 1.1.23; Bates et al., 2014) and *brms* (Bürkner, 2017) packages in R (R Core Team, 2020). Of primary interest are the outcomes of the Bayesian models, particularly because Bayesian approaches allow us to accept the null hypothesis. However, because we created a sampling plan with frequentist techniques, we wanted to include frequentist models as well. Furthermore, given the widespread use of frequentist models in psychological science, including these analyses increases the future replicability of our work. The vibrotactile and noun judgement tasks use response time (in ms) as the outcome, whereas the recall task uses accuracy (0 or 1) as the outcome. We will assume a Gaussian distribution and an identity link function for response time data. Although response times tend to be skewed, linear mixed effects models tend to be robust to violations of the normality assumption, and we will trim extreme response time values before beginning data analysis (see Exclusion Criteria above). For recall data, we will assume a Bernoulli distribution with a logit link function due to the binary nature of the data (i.e., the outcome of each trial is either 0 or 1).

Participants and words will be included as random effects, and following the recommendations of Barr and colleagues (2013) and Brown (2021), we will attempt to fit the model with the maximal theoretically motivated random effects structure justified by the design. Specifically, we will include by-participant and by-item random intercepts, as well as by-participant and by-item random slopes for noise type for all tasks except the noun judgment task. In this task, responses were only included in the analysis if the word was classified as a noun. Given that some words may never or very rarely be classified as nouns, some levels of the random effect for words will have very few observations. This sort of imbalance in the data can cause convergence issues, and random slopes estimates for levels with only a few observations are uninformative, so we will not include by-word random slopes in any of the analyses for the noun judgment task.

If there are estimation issues when fitting frequentist models with the full random effects structure (i.e., convergence or singular fit warnings, perfect correlations among random effects, or variance estimates of exactly 0; Brown, 2021), even when we include control parameters (such as changing the optimizer) or set the correlations among random intercepts and slopes to 0, we will remove the by-item random slope for noise type in the vibrotactile and running memory tasks. Participants tend to vary more than items, and it is conceivable that the effect of noise type is consistent across words. However, we will not remove the by-item random intercepts, as we are assuming item-level differences in response times and recall. Additionally, we will not remove the by-participant random slope for noise type because not including random slopes would amount to assuming that participants do not differ in the extent to which they are affected by the type of background noise, which is an unreasonable assumption to make. In the unlikely event that every frequentist model encounters an estimation issue, we will solely analyze the data with Bayesian multilevel modeling, as these models rarely encounter convergence issues, even when they employ rich random effects structures.

Each Bayesian model will have four MCMC chains of 6,000 iterations total, 2,000 of which are for warm-up. This will result in 16,000 post-warm up samples per model. Divergent transitions during sampling can lead to inaccurate posterior sampling, so if we encounter divergent transitions, we will increase “adapt delta” (a sampler control parameter) from the default value of 0.85 to 0.90. This will slow down the sampler but ensure more accurate transitions. If divergent transitions are still identified, we will continue to increase the “adapt delta” value until there are no longer divergent transitions. To ensure that estimation went smoothly and no convergence issues were encountered, we will check that all R-hat values are equal to 1.00 after sampling. If we observe R-hat values greater than 1.10, suggesting improper convergence (Bürkner, 2017), we will increase the number of iterations by 2,000 and re-run the model (repeating this process if necessary). We will use default priors for all models.

Critical tests. We will assess the effects of background noise on response time or recall separately for each task. For our frequentist analysis, we will compare nested models, one with noise type as a fixed effect and one without, via a likelihood ratio test. If the effect of noise is significant (i.e., the full model provides a better fit for the data than the reduced model), we will examine the summary output of the full model to obtain parameter estimates and to examine the extent to which each of the two noise conditions differ from the reference level (according to a dummy coding scheme in which the silence condition serves as the reference level). Although we do not have a specific prediction about whether listening effort will differ for two-talker babble and speech-shaped noise, we will conduct this pairwise comparison as well. We will use an alpha level of 0.05 for all model comparisons, and we will report all fixed effects estimates with 95% confidence intervals. We will be conducting Bayesian analyses by examining the 95% highest density intervals (HDIs). To do so, we will build a single model with noise as a fixed effect. If the effect of noise type is significant, the lower bound for the 95% HDI will be above zero; if the effect of noise is not significant, the 95% HDI will contain zero (Nicenboim & Vasishth, 2016).

Finding no effects of noise in Experiment 1 would contradict the robustly established finding that noise affects secondary task performance in listening effort paradigms. **Hypothesis 1** states that the model that contains noise as a fixed effect will provide a better fit for the data than the model without it, and that the summary output for the full model will reveal that both the speech-shaped noise and two-talker babble slowed response times and hindered recall relative to silence. This pattern of data would replicate prior work showing that noise increases listening effort, and would help rule out the possibility that any null effects observed in Experiment 2 are attributable to methodological limitations. However, it is possible that the listening conditions we initially choose will be too easy. After collecting data from 20 participants, we will calculate the means for each condition in each task (without performing any further analyses). We will set a minimum threshold of a 30 ms increase in response time and a decrease in recall probability of 0.03 between conditions. These are effect sizes one-third as large as those observed in Brown and Strand (2019). If reaction times are numerically at least 30 ms slower and recall accuracy is numerically at least three percentage points poorer in noise than in silence, we will continue data collection. Given that this finding has been robustly demonstrated many times in previous research, if we do not see a difference between quiet and noise or we see the opposite pattern of results, this would indicate that the noise level we selected was not difficult enough or there was another issue with how the experiment was implemented. In this case, we will terminate data collection, select a more difficult SNR (decreasing by 2 dB), and restart data collection, repeating this process until we obtain the expected detrimental effects of noise in Experiment 1.

Critical tests (NASA-TLX). The primary goal of the proposed experiments is to assess whether the three tasks described above that are commonly used to assess listening effort are also sensitive to changes in background noise in the absence of speech. However, as a supplemental exploratory measure, we will assess the effect of background noise on subjective effort reported for each task separately. We will use linear mixed effects models (assuming a Gaussian distribution with an identity link function) with random intercepts for participants and by-participant random slopes for noise. Random effects for items will not be included in this analysis because participants will respond to a block of stimuli rather than a particular item, and only one item on the NASA-TLX will be analyzed. Given that each participant responded to the effort question four times per noise condition in each task, we will analyze 12 responses per participant for each task. Model fitting and comparison criteria for all subjective effort analyses will be identical to those for our listening effort analyses.

Experiment 2: Orthographic Presentation

Experiment 2 will follow the conventions of Experiment 1, but rather than presenting target words aurally, we will present them orthographically on the screen for participants to read. Any deviations from Experiment 1 are explicitly addressed below.

Participants

We will collect data from 160 participants who did not participate in Experiment 1.

Sample Size Justification

We will collect data from 160 participants for Experiment 2. This sample size is considerably larger than the sample size in Experiment 1 to account for the possibility that the effects in Experiment 2 are smaller than those in Experiment 1. We performed a power analysis using the same procedure and effect sizes as in Experiment 1, except we simulated a sample size of 160 participants. Simulated power was 90% for effect sizes roughly half the size of the minimum effect sizes determined by the power analysis for Experiment 1 (see Figure 3).

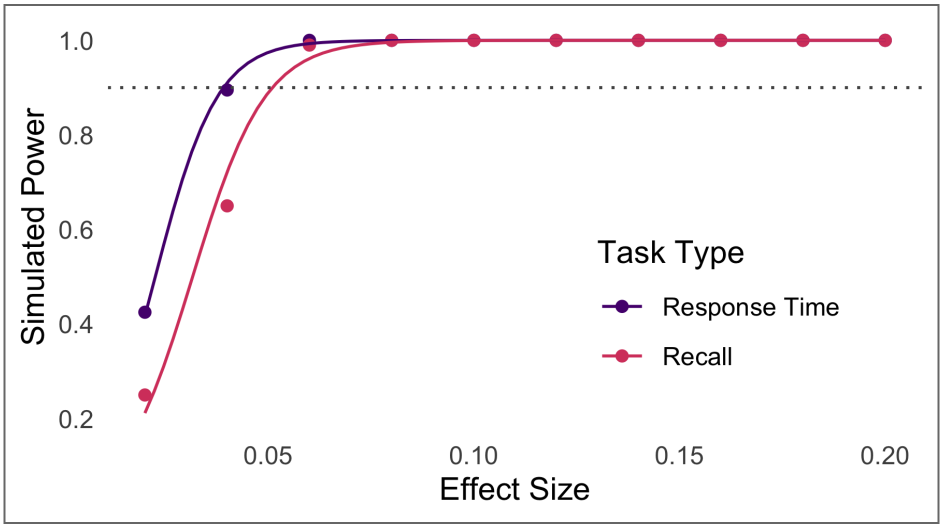

Figure 3. Power curve for Experiment 2. Curve was obtained from simulations with *simr* and data from Brown and Strand (2019a), adjusted to 160 participants. The gray dotted line denotes power of 0.90.

Procedure

Orthographic stimuli will be presented in uppercase black text at the center of a grey screen in a sans serif font. The duration that words will be presented on the screen will match the average duration of the target speech stimuli, and the interstimulus intervals will match those in Experiment 1.

Exclusion Criteria

Exclusion criteria are identical to those in Experiment 1.

Data Analysis

The analyses of Experiment 2 will mirror those in Experiment 1, but given that we have different predictions about how steady-state noise and two-talker babble affect task performance, we will compare each noise type to silence in separate analyses (i.e., noise type had two levels rather than three). Recall that **Hypothesis 2a** states that speech-shaped noise will not impair performance on the tasks relative to silence, and **Hypothesis 2b** states that the presence of two-talker babble will negatively affect all tasks relative to silence. Thus, given these specific hypotheses about how the two types of noise affect performance on each task, we plan to conduct two separate analyses in which we directly compare speech-shaped noise to silence (to assess **Hypothesis 2a**) and two-talker babble to silence (to assess **Hypothesis 2b**). That is, we will build separate models corresponding to **Hypothesis 2a** and **Hypothesis 2b** in which noise type has only two levels (either silence and steady-state noise or silence and two-talker babble). As in the analyses for Experiment 1, we will use nested model comparisons for frequentist models and 95% highest density intervals for Bayesian models to assess whether a model that contains noise provides a better fit for the data than a model without it. Should we obtain a null result in our frequentist model comparisons, we will perform equivalence testing (see Lakens, 2017).

Finally, in Experiment 2 we expect that two-talker babble (but not speech-shaped noise) would impair performance on all three tasks, but **Hypothesis 2c** further states that the babble interference effect will be largest for the running memory task, followed by the noun judgment task, and then by the vibrotactile task. It is not possible to statistically compare performance on the three tasks in a single model because recall outcomes are binary and dual-task outcomes are continuous. However, after performing statistical analyses on each task separately, we will compare the magnitude of the effect size (Cohen's *d*) across the three tasks, where the effect size refers to the difference between the silence and two-talker babble conditions (see Johnson et al., 2015; Picou & Ricketts, 2014; Strand et al., 2018 for comparisons of effect sizes across listening effort tasks).

The NASA-TLX analyses will be identical to those described in Experiment 1.

Comparison Across Experiments

It is possible that noise-induced impairments on listening effort tasks are the result of both listening effort from processing degraded speech and more general noise-induced cognitive interference. In that case, performance impairments should be larger in Experiment 1 than Experiment 2, as Experiment 2 cannot induce listening effort associated with processing degraded speech. We will not analyze the data from Experiments 1 and 2 in the same model because the time course of hearing and reading speech are quite different, so response times between experiments are expected to differ for reasons other than effects of background noise. However, we will again calculate effect sizes (Cohen's *d*) to assess the extent to which each noise type impaired performance relative to silence for each task in Experiment 1 compared to Experiment 2. The magnitude of the difference between the effect sizes (referring to the difference between the silence and steady-state noise conditions as well as the difference between the silence and two-talker babble conditions, in each task separately) for Experiment 1 and Experiment 2 provides an indication of the extent to which findings typically ascribed to listening effort were in fact due to listening effort from processing degraded speech rather than to noise-induced cognitive interference.

References

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

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A., & Salamé, P. (1986). The unattended speech effect: perception or memory? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *12*(4), 525–529.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*(3), 445–459.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., & Green, P. (2014). *Package "lme4"* (Version 1.1-15). R foundation for statistical computing, Vienna, 12. https://github.com/lme4/lme4/

Beaman, C. P., & Jones, D. M. (1997). Role of serial order in the irrelevant speech effect: Tests of the changing-state hypothesis. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *23*(2), 459.

Beaman, C. P., Philip Beaman, C., & Jones, D. M. (1998). Irrelevant Sound Disrupts Order Information in Free Recall as in Serial Recall. In *The Quarterly Journal of Experimental Psychology Section A* (Vol. 51, Issue 3, pp. 615–636). https://doi.org/10.1080/713755774

Bench, J., Kowal, A., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, *13*(3), 108–112.

Borghini, G., & Hazan, V. (2018). Listening Effort During Sentence Processing Is Increased for

Non-native Listeners: A Pupillometry Study. *Frontiers in Neuroscience*, *12*, 152.

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: language familiarity and semantic content. *The Journal of the Acoustical Society of America*, *131*(2), 1449–1464.

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920960351.

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. *The Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/1747021820916726

Brown, V. A., & Strand, J. F. (2018). Noise increases listening effort in normal-hearing young adults, regardless of working memory capacity. *Language, Cognition and Neuroscience*, *34*, 628–640.

Brown, V. A., & Strand, J. F. (2019a). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition*, *2*(1). https://doi.org/10.5334/joc.89

Brown, V. A., & Strand, J. F. (2019b). “Paying” attention to audiovisual speech: Do incongruent stimuli incur greater costs? *Attention, Perception & Psychophysics*, *81*(6), 1743–1756.

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997.

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, Articles*, *80*(1), 1–28.

Campbell, T., Beaman, C. P., & Berry, D. C. (2002). Changing-state disruption of lip-reading by irrelevant sound in perceptual and memory tasks. *The European Journal of Cognitive Psychology*, *14*(4), 461–474.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning and Verbal Behavior*, *15*(1), 17–31.

Dobbs, S., Furnham, A., & McClelland, A. (2011). The effect of background music and noise on the cognitive test performance of introverts and extraverts. *Applied Cognitive Psychology*, *25*(2), 307–313.

Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research: JSLHR*, *53*(1), 18–33.

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588.

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

Gosselin, P. A., & Gagné, J.-P. (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *International Journal of Audiology*, *50*(11), 786–792.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183.

Johnson, J., Xu, J., Cox, R., & Pendergraft, P. (2015). A comparison of two methods for measuring listening effort as part of an audiologic test battery. *American Journal of Audiology*, *24*(3), 419–431.

Jones, D. M., Alford, D., Bridges, A., Tremblay, S., & Macken, B. (1999). Organizational factors in selective attention: The interplay of acoustic distinctiveness and auditory streaming in the irrelevant sound effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*(2), 464–473.

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. In *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 19, Issue 2, pp. 369–381). https://doi.org/10.1037//0278-7393.19.2.369

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory? In *Applied Cognitive Psychology* (Vol. 4, Issue 2, pp. 89–108). https://doi.org/10.1002/acp.2350040203

Jones, D. M., & Morris, N. (1992). Irrelevant speech and serial recall: implications for theories of attention and working memory. *Scandinavian Journal of Psychology*, *33*(3), 212–229.

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Lidestam, B., Holgersson, J., & Moradi, S. (2014). Comparison of informational vs. energetic masking effects on speechreading performance. *Frontiers in Psychology*, *5*, 639.

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, *58*(1), 22–33.

McLaughlin, D. J., & Van Engen, K. J. (2020). Task-evoked pupillary response for intelligible accented speech. *The Journal of the Acoustical Society of America*, *147*. https://doi.org/10.1121/10.0000718

Morey, R. D., Kaschak, M. P., Díez-Álamo, A. M., Glenberg, A. M., Zwaan, R. A., Lakens, D., Ibáñez, A., García, A., Gianelli, C., Jones, J. L., & Others. (2021). A pre-registered, multi-lab non-replication of the action-sentence compatibility effect (ACE). *Psychonomic Bulletin & Review*. https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_3309331

Morris, N., & Jones, D. M. (1990). Memory updating in working memory: The role of the central executive. *British Journal of Psychology* , *81*(2), 111–121.

Myerson, J., Spehar, B., Tye-Murray, N., Van Engen, K., Hale, S., & Sommers, M. S. (2016). Cross-modal informational masking of lipreading by babble. *Attention, Perception & Psychophysics*, *78*(1), 346–354.

Neath, I. (2000). Modeling the effects of irrelevant speech on memory. *Psychonomic Bulletin & Review*, *7*(3), 403–423.

Norris, D., Baddeley, A. D., & Page, M. P. A. (2004). Retroactive effects of irrelevant speech on serial recall from short-term memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(5), 1093–1105.

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 Suppl 1*, 5S – 27S.

Pichora-Fuller, M. K., Schneider, B. A., & Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *The Journal of the Acoustical Society of America*, *97*(1), 593–608.

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.

Pisoni, D. B. (1996). Word Identification in Noise. *Language and Cognitive Processes*, *11*(6), 681–687.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology*, *20*(3), 241–248.

R Core Team. (2020). *R 4.0.2*. R Foundation for Statistical Computing Vienna, Austria.

Rennies, J., Schepker, H., Holube, I., & Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *The Journal of the Acoustical Society of*

America, *136*(5), 2642–2653.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, O., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience*, *7*, 31.

Salamé, P., & Baddeley, A. (1982). Disruption of short-term memory by unattended speech: Implications for the structure of working memory. *Journal of Verbal Learning and Verbal Behavior*, *21*(2), 150–164.

Salamé, P., & Baddeley, A. (1987). Noise, unattended speech and short-term memory. *Ergonomics*, *30*(8), 1185–1194.

Salamé, P., & Baddeley, A. (1989). Effects of Background Music on Phonological Short-Term Memory. In *The Quarterly Journal of Experimental Psychology Section A* (Vol. 41, Issue 1, pp. 107–122). https://doi.org/10.1080/14640748908402355

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: JSLHR*, *52*(5), 1230–1240.

Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research: JSLHR*, *58*(6), 1781–1792.

Sommers, M. S., & Phelps, D. (2016). Listening effort in younger and older adults: A comparison of auditory-only and auditory-visual presentations. *Ear and Hearing*, *37 Suppl 1*, 62S – 8S.

Sommers, M. S., Tye-Murray, N., Barcroft, J., & Spehar, B. P. (2015). The effects of meaning-based auditory training on behavioral measures of perceptual effort in individuals with impaired hearing. *Seminars in Hearing*, *36*(4), 263–272.

Strand, J. F., Brown, V. A., & Barbour, D. L. (2020). Talking points: A modulating circle increases listening effort without improving speech recognition in young adults. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01713-y

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: JSLHR*, *61*, 1463–1486.

Szalma, J. L., & Hancock, P. A. (2011). Noise effects on human performance: a meta-analytic synthesis. *Psychological Bulletin*, *137*(4), 682–707.

Tremblay, S., MacKen, W. J., & Jones, D. M. (2001). The impact of broadband noise on serial memory: Changes in band-pass frequency increase disruption. *Memory* , *9*(4), 323–331.

Van Engen, K. J. (2010). Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble. *Speech Communication*, *52*(11-12), 943–953.

Van Engen, K. J., & McLaughlin, D. J. (2018). Eyes and ears: Using eye tracking and pupillometry to understand challenges to speech recognition. *Hearing Research*, *369*, 56–66.

Wais, P. E., & Gazzaley, A. (2011). The impact of auditory distraction on retrieval of visual memories. *Psychonomic Bulletin & Review*, *18*(6), 1090–1097.

Weisz, N., & Schlittmeier, S. J. (2006). Detrimental effects of irrelevant speech on serial recall of visual items are reflected in reduced visual N1 and reduced theta activity. *Cerebral Cortex* , *16*(8), 1097–1105.

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*. https://doi.org/10.1177/2331216516669723

Winn, M. B. (2018). *Praat script for creating speech-shaped noise [software] version 12*. http://www.mattwinn.com/praat.html