JSLHR

Research Article

Measuring Listening Effort: Convergent Validity, Sensitivity, and Links With Cognitive and Personality Measures

Julia F. Strand,[a] Violet A. Brown,[a] Madeleine B. Merchant,[a] Hunter E. Brown,[a] and Julia Smith[a]

Purpose: Listening effort (LE) describes the attentional or cognitive requirements for successful listening. Despite substantial theoretical and clinical interest in LE, inconsistent operationalization makes it difficult to make generalizations across studies. The aims of this large-scale validation study were to evaluate the convergent validity and sensitivity of commonly used measures of LE and assess how scores on those tasks relate to cognitive and personality variables.

Method: Young adults with normal hearing (N = 111) completed 7 tasks designed to measure LE, 5 tests of cognitive ability, and 2 personality measures.

Results: Scores on some behavioral LE tasks were moderately intercorrelated but were generally not correlated with subjective and physiological measures of LE, suggesting that these tasks may not be tapping into the same underlying construct. LE measures differed in their sensitivity to changes in signal-to-noise ratio and the extent to which they correlated with cognitive and personality variables.

Conclusions: Given that LE measures do not show consistent, strong intercorrelations and differ in their relationships with cognitive and personality predictors, these findings suggest caution in generalizing across studies that use different measures of LE. The results also indicate that people with greater cognitive ability appear to use their resources more efficiently, thereby diminishing the detrimental effects associated with increased background noise during language processing.

Processing spoken language requires extracting sensory information from a rapidly changing acoustic signal and making a series of perceptual and cognitive judgments. The difficulty of this task varies depending on a host of factors, including presence and type of background noise, content of the speech, characteristics of the listener, signal degradation, speaker characteristics, and many others (see Mattys, Davis, Bradlow, & Scott, 2012, for a review). More challenging conditions are likely to require greater listening effort (LE), the attentional or cognitive requirements for successful listening (cf., Bourland-Hicks & Tharpe, 2002; Downs, 1982). The concept of LE has also been defined in terms of the "mental exertion" (McGarrigle et al., 2014) or "mental effort" (Pichora-Fuller et al., 2016) required during a listening task, but all conceptualizations of LE draw attention to the high-level cognitive (as opposed to low-level sensory) aspects of listening. A crucial assumption of LE research is that the cognitive system has a limited pool of resources (Kahneman, 1973; Pashler, 1994), so as speech becomes more difficult to parse and requires more effort to recognize, there are fewer resources remaining to devote to other tasks (Pichora-Fuller et al., 2016; Wingfield, 2016). Prior research has indicated that greater LE is required for listeners to understand speech in noisy rather than quiet conditions (e.g., Downs & Crum, 1978; Rudner, Lunner, Behrens, Thorén, & Rönnberg, 2012), for people with hearing impairments relative to those with normal hearing (Bourland-Hicks & Tharpe, 2002), for dichotic (different digits presented to each ear) compared to diotic (the same digit presented to both ears) speech (Seeman & Sims, 2015), for competing speech relative to stationary noise (Ng, Rudner, Lunner, Pedersen, & Rönnberg, 2013), and for situations in which the location of the signal varies rather than being constant (Koelewijn, de Kluiver, Shinn-Cunningham, Zekveld, & Kramer, 2015).

There is general agreement about many aspects of LE, including that effort can be affected by the physical listening situation (i.e., noise type and level; Downs & Crum, 1978; Rudner et al., 2012), speech style (Van Engen, Chandrasekaran, & Smiljanic, 2012), speech content (Johnson, Xu, Cox, & Pendergraft, 2015), and participant

[a]Department of Psychology, Carleton College, Northfield, MN

Correspondence to Julia Strand: jstrand@carleton.edu

Editor-in-Chief: Frederick (Erick) Gallun
Editor: Lori J. Leibold

Received July 7, 2017
Revision received November 21, 2017
Accepted February 6, 2018
https://doi.org/10.1044/2018_JSLHR-H-17-0257

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

characteristics (Gosselin & Gagné, 2011a, 2011b). In addition, there is widespread consensus that maintaining high levels of LE can have negative consequences for listeners, including subjective reports of mental fatigue (Alhanbali, Dawes, Lloyd, & Munro, 2017; McGarrigle, Dawes, Stewart, Kuchinsky, & Munro, 2016) and distress (Kramer, Kapteyn, & Houtgast, 2006). The construct of LE also has been of interest to clinicians given its applications for treating people with hearing loss (cf., Alhanbali et al., 2017; Desjardins & Doherty, 2014). For example, noise reduction algorithms in hearing aids may reduce the effort necessary to understand speech, even if they do not improve recognition accuracy (Desjardins & Doherty, 2014; Sarampalis, Kalluri, Edwards, & Hafter, 2009). Thus, assessing the consequences of noise reduction algorithms by relying solely on speech intelligibility measures may provide an incomplete picture about a patient's experience with a hearing aid.

Despite substantial theoretical and clinical interest in LE, it has been described as a "poorly determined" (Rudner, Ng, et al., 2011, p. 47) construct. The *International Journal of Audiology* even published a discussion paper entitled "Listening effort and fatigue: What exactly are we measuring?" (McGarrigle et al., 2014). More recently, however, researchers have worked toward clarifying and operationalizing LE (see Pichora-Fuller et al., 2016), but a remaining impediment to a full understanding of LE is the lack of consensus on how to measure and quantify it. Our survey of prior research revealed more than two dozen different tests in the literature that are intended to measure LE, yet there has been little psychometric work evaluating whether and how these measures are related to one another. Perhaps as a result of this measurement variability, there are numerous unresolved, contradictory findings in the literature, including whether individual differences in the amount of LE expended are related to differences in cognitive abilities such as working memory (WM; for instance, see Desjardins & Doherty, 2014; Ng et al., 2013), and whether seeing and hearing a talker increases or decreases the amount of LE that is required, relative to hearing alone (see Mishra, Lunner, Stenfelt, Rönnberg, & Rudner, 2013a, 2013b; Sommers & Phelps, 2016).

Although a large body of research has sought to explain how listening conditions and participant characteristics affect LE, the lack of psychometric evaluation of LE tasks (i.e., tasks that are expected to be measuring LE) leaves three important questions unanswered. First, are these multiple measures of LE in fact tapping into the same underlying construct? If the effort necessary to understand speech in consistent conditions is relatively stable within individuals, participants who perform well on one LE task may be expected to do well on another. Thus, observing strong correlations among multiple measures of LE would suggest that the tests are measuring a single faculty on which individuals reliably differ. Second, how do individual measures of LE differ in their sensitivity to changes in task difficulty? Even if multiple measures of LE are tapping into the same underlying construct, some may be more sensitive than others at identifying subtle changes in task difficulty. This information would be particularly useful to clinicians seeking to effectively quantify LE in short clinical visits. Finally, are individual differences in performance on tasks that require LE related to individual differences in other cognitive abilities? A recent consensus paper conceptualized LE as a specific instance of mental effort—"the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a task"—when the task involves listening to speech or other auditory inputs (Pichora-Fuller et al., 2016, p. 10S). If LE is indeed a form of mental effort, it follows that the amount of effort expended during a listening task may be related to a listener's ability to utilize other cognitive resources. Given that many behavioral measures of LE evaluate the amount of cognitive resources available after successful listening has occurred (Mishra et al., 2013b; Picou, Ricketts, & Hornsby, 2011; Rudner, Ng, et al., 2011; Sommers & Phelps, 2016), it might be expected that an individual with greater general cognitive capacity (e.g., executive control, WM capacity, processing speed) would have an advantage on tasks that are assumed to measure LE because they use their resources more efficiently. In contrast, making use of those resources may facilitate speech understanding at the cost of an increase in processing load (see Zekveld, Kramer, & Festen, 2011).

The current project seeks to address these gaps in the literature. Specifically, the goals of this study are to (a) evaluate relationships among multiple existing measures of LE to assess the convergent validity of these measures, (b) compare the sensitivity of LE measures to a consistent change in signal-to-noise ratio (SNR), and (c) assess whether measures of LE rely on well-established measures of individual differences in cognition. In the next section, we review the broad categories of LE measures that have been used previously and justify the measures we opted to include in the current study. Then, we describe prior work that is relevant to each of our three study goals.

Measures of LE

Measures of LE that have been used in the literature can broadly be broken down into three categories: subjective, behavioral, and physiological (for additional review of measures of LE, see Gagné, Besser, & Lemke, 2017; McGarrigle et al., 2014; Pichora-Fuller et al., 2016). See Table 1 for a summary of the measures used in the current study.

Subjective Measures

Subjective self-report measures of LE are widely used and typically involve participants reporting how much effort they felt they had to expend to successfully complete the listening task. These measures have included rating the degree of perceived effort during performance on the tasks using a 1–7 scale (Johnson et al., 2015), a 0–10 scale (Koelewijn et al., 2015), or the Borg CR-10 scale (which allows participants to report higher than 10 if they feel it is necessary; Borg, 1990; Hällgren, Larsby, Lyxell, & Arlinger,

Table 1. Summary of all tasks.

Category	Task	Description and representative citation
LE: subjective	NASA Task Load Index (NASA-TLX)	Subjective ratings of mental demand, perceived performance, effort exerted, and frustration (Seeman & Sims, 2015).
LE: behavioral (reaction time)	Complex dual-task (CDT)	Hear and repeat words and respond to visually presented odd and even numbers (Sarampalis et al., 2009).
	Semantic dual-task (SDT)	Hear and repeat words and judge whether each word is a noun (Picou & Ricketts, 2014).
LE: behavioral (recall)	Cognitive Spare Capacity Test (CSCT)	Hear numbers produced by male and female talkers, and keep track of highest/lowest or odd/even (Mishra et al., 2013a, 2013b).
	Listening span (LS)	Hear sentences, repeat final word, make predictability judgment, recall all final words at the end of a series (Pichora-Fuller et al., 1995).
	Running memory (RM)	Hear lists of words, recall last three at the end of a series (Sommers & Phelps, 2016).
LE: physiological	Pupillometry	Hear and repeat sentences while pupil size is monitored (Zekveld & Kramer, 2014).
Audiological	Audiogram	Pure-tone thresholds at 250, 500, 1000, 2000, 4000, and 8000 Hz for both ears.
Cognitive	Reading span (Rspan; WM)	Read sentences, repeat final word, make sensicality judgment, recall all final words at the end of a series (Daneman & Carpenter, 1980).
	Simon task (Simon; inhibition)	Respond to red and blue rectangles via right and left key presses (Mishra et al., 2013a, 2013b).
	Letter memory (LM; updating)	See strings of letters, recall last four (Mishra et al., 2013a, 2013b).
	Lexical decision task (LDT; processing speed)	Make word or nonword judgments on letter strings (Picou et al., 2013).
	Text Reception Threshold (TRT) test (linguistic closure)	Read masked sentences (Besser et al., 2012).
Personality	Big Five Inventory–2 (BFI-2; extraversion)	Self-report personality questionnaire (Soto & John, 2017) to measure extraversion.
	Highly Sensitive Person Scale (HSPS; sensitivity)	Self-report sensitivity questionnaire (Aron & Aron, 1997).

Note. LE tasks were completed in easy and hard conditions. LE = listening effort; NASA = National Aeronautics and Space Administration; WM = working memory.

2005; Larsby, Hällgren, Lyxell, & Arlinger, 2005). Participants may also be asked to estimate the percentage of sentences identified correctly (Fraser, Gagné, Alepins, & Dubois, 2010; Gosselin & Gagné, 2011b); evaluate mental demand, effort, performance (error rate), and frustration using the NASA Task Load Index (NASA-TLX) 21-point scale (Hart & Staveland, 1988; Seeman & Sims, 2015); or report on motivation (Koelewijn, Zekveld, Festen, Rönnberg, & Kramer, 2012). Participants typically report greater perceived effort as SNR decreases (e.g., Larsby et al., 2005; Seeman & Sims, 2015; Zekveld, Kramer, & Festen, 2010), for audio-only relative to audiovisual presentations (Fraser et al., 2010), and when the location of the signal is varied relative to when it is consistent (Koelewijn et al., 2015).

In the current study, self-report measures of LE were collected multiple times during each behavioral and physiological task using the NASA-TLX, which provides participants the opportunity to differentiate between how successfully they believe they completed the task (their performance) and the effort required to do so (Hart & Staveland, 1988; Seeman & Sims, 2015).

Behavioral Measures

Dual-task paradigms. A commonly used type of behavioral LE measure involves assessing performance on a secondary task that is administered while participants simultaneously complete a speech recognition task (see Gagné et al., 2017, for a recent review). Participants are typically told to focus their attention on the primary word or sentence recognition task, with the assumption being that good performance on that task requires enough cognitive resources that the remaining resources are insufficient to efficiently perform the secondary task, resulting in detriments in performance on the secondary task (Bourland-Hicks & Tharpe, 2002; Desjardins & Doherty, 2013; Downs, 1982; Fraser et al., 2010). Broadbent (1958) was the first to demonstrate the consequences of LE behaviorally by showing that listeners performed more poorly on a secondary visual distractor task when the primary speech task became more difficult. This pioneering work indicated that equivalent word recognition accuracies do not necessarily indicate equivalent cognitive effort exerted to process speech. Since then, dual-task costs have been demonstrated with other tasks, such as memorizing digits (Howard, Munro, & Plack, 2010; Rakerd, Seitz, & Whearty, 1996), responding to the appearance of a probe light (Bourland-Hicks & Tharpe, 2002; Downs, 1982; Downs & Crum, 1978), making speeded judgments about visual stimuli (Sarampalis et al., 2009; Seeman & Sims, 2015), responding to vibrotactile patterns (Fraser et al., 2010; Gosselin & Gagné, 2011a, 2011b), performing visual matching (Hughes & Galvin, 2013), tracking visual objects (Desjardins & Doherty, 2013; Tun, McCoy,

& Wingfield, 2009), doing a mental rotation task (Pals, Sarampalis, & Baskent, 2013), completing a simulated driving task in which participants must follow a car that varies its velocity without crashing (Wu et al., 2014), inhibiting irrelevant information using the Stroop task (Wu, Stangl, Zhang, Perkins, & Eilers, 2016), completing a dot-to-dot game (Pittman, 2011), and making semantic or rhyme judgments (Pals et al., 2013; Picou & Ricketts, 2014), among others. Although the particulars of the tasks vary, slower and less accurate performance on the secondary task is typically observed in conditions that are assumed to require greater LE (e.g., more difficult SNRs).

Given the prevalence of dual-task measures in the LE literature, we opted to include two such paradigms in the current study: the complex dual-task (CDT; Picou & Ricketts, 2014; Sarampalis et al., 2009), in which participants make speeded judgments about whether visually presented numbers are even or odd while simultaneously completing a word recognition task, and the semantic dual-task (SDT; Picou & Ricketts, 2014), which requires making a speeded judgment about whether each word in an aurally presented list is a noun. Picou and Ricketts (2014) argue that the SDT requires greater depth of processing, so it taxes cognitive resources more than the CDT. The two tasks also differ in the modality in which the distractor is presented, which may influence the extent to which the primary and secondary tasks compete for processing resources in young adults (Guerreiro, Murphy, & Van Gerven, 2013). Including both the CDT and SDT in the current study enables us to evaluate the costs associated with performing a dual-task generally and compare the sensitivity of a shallow, visual task (the CDT) to that of a deeper, verbal one (the SDT).

Recall paradigms. Researchers have also used a variety of memory tasks to measure LE, with the rationale that as LE increases, there will be fewer resources available to encode speech material into memory. An early demonstration of this came from Rabbitt (1968), who showed poorer recall for the initial items of a list when later items were presented in noise than in quiet. Variations on recall tasks have included the running memory (RM) task (McCoy et al., 2005; Sommers & Phelps, 2016), paired-associates task (Murphy, Craik, Li, & Schneider, 2000; Picou et al., 2011), and tasks involving repeating and then recalling the final words of a series of sentences (sometimes referred to as the listening span [LS] task; Johnson et al., 2015; Ng et al., 2013; Pichora-Fuller, Schneider, & Daneman, 1995; Sarampalis et al., 2009; see below for more on the LS task in the WM and LE literatures).

A recently introduced measure called the Cognitive Spare Capacity Test (CSCT; Keidser, Best, Freeston, & Boyce, 2015; Mishra et al., 2013a, 2013b; Mishra, Stenfelt, Lunner, Rönnberg, & Rudner, 2014; Rudner, Ng, et al., 2011) was designed to measure the "residual cognitive capacity once successful listening has taken place" (Rudner, Ng, et al., 2011, p. 47). In this task, participants listen to numbers spoken by a male and a female voice and are asked to monitor the highest or lowest number spoken by each talker (updating condition), or to recall the odd or even numbers spoken by a single specified talker (inhibition condition) in low-load (recalling two items) and high-load (recalling three items, including the first number) conditions.

Although the implementation of recall paradigms differs, in all cases, poorer recall is expected as LE increases. In the current study, we administered the RM task, which involves presentation of individual words; the LS task, which includes high-predictability (HP) and low-predictability (LP) sentences; and the CSCT, which relies on separable components of updating and inhibiting in both high-load and low-load conditions. These paradigms were selected because they are frequently used in the literature, but the types of stimuli used and instructions differ, so they may be expected to place different demands on listeners. As one example, the RM and LS tasks both use words and sentences (rather than numbers) as stimuli, but the RM task is primarily an updating task in which participants must keep track of the most recent words but do not need to process them deeply, whereas the LS task additionally requires making judgments about the predictability of the sentence. Thus, these tasks may be expected to put different demands on listeners' WM.

Physiological Measures

Given the well-established link between cognitive demands and physical changes in the body (e.g., Kahneman & Beatty, 1966), increases in the effort expended to understand speech may be reflected in physiological markers of stress and arousal. For example, increases in task demands result in higher levels of activation in the sympathetic nervous system, which causes increases in electromyographic activity in facial muscles (Mackersie & Cones, 2011) and changes in heart rate variability (Seeman & Sims, 2015). Participants also tend to show increased skin conductance as task demands increase (Mackersie & Cones, 2011; Seeman & Sims, 2015). However, skin conductance appears not to be sensitive to changes in SNR (Seeman & Sims, 2015), a manipulation that has been robustly shown to affect LE using other measures (Downs & Crum, 1978; Rudner et al., 2012). Neuroanatomical markers, including activity in the prefrontal cortex, premotor cortex, and the cingulo-opercular network areas, have also been associated with LE (see Peelle, 2017 for an excellent review of this literature).

The most commonly used physiological measure of LE is pupil dilation; when a task requires more effort, average pupil dilation is larger (Beatty, 1982). Increased pupil dilation in difficult listening conditions results from activity in the locus coeruleus, a major noradrenergic nucleus in the brain that is associated with stress and arousal (Alnæs et al., 2014). Pupil dilation is sensitive to changes in SNR (Kramer, Kapteyn, Festen, & Kuik, 1997; Zekveld, Heslenfeld, Johnsrude, Versfeld, & Kramer, 2014; Zekveld & Kramer, 2014; Zekveld et al., 2010), vocoded signal degradation (Winn, Edwards, & Litovsky, 2015), single-talker masking compared with other types of masking (Koelewijn, Zekveld, Festen, & Kramer, 2012), lexical competition (Kuchinsky et al., 2013; Wagner, Toffanin, & Başkent, 2016), syntactic complexity (Piquado, Isaacowitz, & Wingfield, 2010; Wendt, Dau, & Hjortkjær, 2016), location of the speech signal (Koelewijn et al., 2015), and divided attention over two speech streams as opposed to focused attention on one speech stream (Koelewijn, Shinn-Cunningham, Zekveld, & Kramer, 2014). We opted to include pupillometry as the physiological measure in the current study given its prevalence in the literature.

Research Aims

Aim 1: Assess the Convergent Validity of LE Measures

The bulk of the psychometric work on LE has compared subjective and objective measures of LE, and this has generally shown that subjective measures are correlated with task accuracy but not with objective measures of effort, such as dual-task performance (Downs & Crum, 1978; Feuerstein, 1992; Gosselin & Gagné, 2011b; Seeman & Sims, 2015), changes in word recall (Johnson et al., 2015), or physiological measures (Seeman & Sims, 2015). Interestingly, Koelewijn, Zekveld, Festen, and Kramer (2012) found that LE as measured by pupillometry was correlated with subjective ratings of performance rather than subjective ratings of effort. Thus, it appears that participants' perceived LE is not predictive of the declines in dual-task or word recall performance or physical changes associated with increased task demands. This provides some evidence against the idea that multiple measures of LE (behavioral, subjective, and physiological) are tapping into a single underlying ability.

Relatively few studies have correlated multiple behavioral measures of LE. Sarampalis et al. (2009) included both a recall task and a dual-task reaction time measure of LE and found that the two tasks showed similar patterns of results (a noise reduction algorithm reduced LE as measured by both metrics). However, that study did not correlate scores across participants to assess whether participants who performed well on one measure of LE also performed well on the other. Mishra et al. (2013b) included both a word recall task and the CSCT to measure LE and found that, in the absence of background noise, seeing and hearing the talker (relative to hearing alone) decreased cognitive spare capacity (CSC) but improved serial recall. Thus, two metrics that have been used to measure LE—CSC and recall—rendered opposite results. Comparisons across studies on audiovisual speech also yield conflicting findings, with work using dual-task paradigms suggesting that processing audiovisual speech (relative to audio-only speech) can increase LE (Fraser et al., 2010; Gosselin & Gagné, 2011a), but research using recall paradigms indicating that audiovisual speech can reduce LE (Sommers & Phelps, 2016). Although the contradictory findings may be attributable to the difficulty of the materials or the conditions of presentation (Sommers & Phelps, 2016), it is also possible that they are the result of using different methods of quantifying LE. Therefore, it is not currently clear whether multiple measures of LE are in fact measuring the same underlying variable, making it difficult to generalize across studies employing different measures of LE.

Aim 2: Assess the Sensitivity of LE Measures

Although there is evidence that manipulating listening difficulty (e.g., changing SNR) affects multiple measures of LE (e.g., subjective effort, dual-task performance), few studies have assessed the sensitivity of multiple measures of LE to methodological manipulations. In one of the most comprehensive comparisons of LE measures to date, Seeman and Sims (2015) compared physiological measures (including heart rate variability and skin conductance), subjective ratings, and a dual-task measure. All the measures were sensitive to changes in either task complexity, SNR, or both. This study also showed that the dual-task measure was more sensitive than the skin conductance measure to changes in task complexity and as sensitive as heart rate variability to changes in SNR. Heart rate variability was more sensitive than skin conductance to changes in task complexity.

Johnson et al. (2015) found that a self-report measure of LE was more sensitive to changes in context (i.e., it showed larger differences between high- and low-predictability sentences) and differences in SNR than a word recall measure. Picou and Ricketts (2014) showed that a secondary task that required deeper processing of verbal material was more sensitive to changes in SNR than a secondary task that required shallower processing. Thus, some measures of LE clearly differ in their sensitivity to changes in task difficulty. Given that it is difficult to make comparisons of effect sizes across studies that include different speakers, materials, participants, and listening conditions, in order to assess the sensitivity of measures of LE, it is necessary to include a broad range of measures in a single study. The current study assessed how changing SNR affects scores on multiple measures of LE.

Aim 3: Evaluate Relationships Between LE and Cognitive Abilities and Personality Traits

Processing spoken language clearly relies on general cognitive mechanisms; even after controlling for audibility, individual differences in cognitive abilities regularly account for significant unique variance in performance on speech tasks (e.g., Humes, 2007; Lunner, 2003; see Besser, Koelewijn, Zekveld, Kramer, & Festen, 2013, for a review). However, it is not clear that improvements in performance on a task correspond to decreases in the effort necessary to complete that task. Ahern and Beatty (1979) originally proposed three frameworks for describing how cognitive capacity may affect processing load, and these were later refined by Van Der Meer et al. (2010; see Zekveld et al., 2011, for more on these hypotheses). Table 2 presents a summary of these hypotheses. The effort hypothesis suggests that people with high cognitive ability invest more of their resources in a task, which leads to better performance on the task at the cost of increased processing load, and this is true regardless of task difficulty. The resource hypothesis

Table 2. Four frameworks and the predictions they make about how cognitive abilities relate to the effort necessary to successfully process speech.

Hypothesis	Effect of high cognitive ability on load	Conditions under which cognitive ability affects load
Effort hypothesis	Increases load	Any
Resource hypothesis	Increases load	Only if difficult
Cognitive efficiency hypothesis	Decreases load	Any
ELU hypothesis	Decreases load	Only if difficult

Note. ELU = ease of language understanding.

also predicts that people with high cognitive ability will show increased processing load, but only in difficult listening conditions. In contrast to hypotheses that greater cognitive capacity leads to increases in processing load, the cognitive efficiency hypothesis predicts that people with high cognitive ability use their resources more efficiently, which reduces processing load, regardless of task difficulty. Another possibility is provided by the Ease-of-Language Understanding (ELU; Rönnberg, 2003; Rönnberg et al., 2013; Rönnberg, Rudner, Foo, & Lunner, 2008) model, which predicts that high cognitive ability reduces processing load specifically in challenging listening situations. That is, when there are mismatches between the acoustic-phonetic input and the representations of words in long-term memory, cognitive resources are recruited to reevaluate the input (see Rönnberg, 2003). These four frameworks make different predictions about whether and how cognitive abilities relate to the effort necessary to successfully process speech.

Cognitive abilities. The cognitive ability that has received the most attention in the LE literature is WM. WM has been shown to be a particularly robust predictor of the ability to recognize speech in noise (Akeroyd, 2008; but see Füllgrabe & Rosen, 2016), but it is less clear whether individual differences in WM also affect the amount of LE an individual must expend. Rudner et al. (2012) showed that higher WM capacity tended to be associated with lower subjectively rated LE. These results suggest that WM may moderate the amount of LE that participants require to understand the speech signal (in line with the predictions of the cognitive efficiency and ELU hypotheses). However, a commonly used WM measure, reading span (Rspan; Daneman & Carpenter, 1980), is not reliably correlated with LE measures, including the CSCT (Mishra et al., 2013b) and a visual-tracking dual-task measure (Desjardins & Doherty, 2014), suggesting that some LE tasks are not affected by WM (but see Mishra et al., 2013a, for a demonstration of a positive correlation between Rspan and CSCT score in quiet).

Interpreting the relationship between LE and WM is complicated by the fact that the LS task (involving recognition and later recall of the final word of sentences) has been independently used both as a measure of WM (Daneman & Carpenter, 1980; Zekveld & Kramer, 2014) and a measure of LE (Johnson et al., 2015; Ng et al., 2013; Pichora-Fuller, Schneider, & Daneman, 1995; Sarampalis et al., 2009). Further, LS is significantly correlated with other well-established measures of WM, such as Rspan (Ng et al., 2013). Thus, it is difficult to draw conclusions about how WM contributes to LE without a better understanding of how to dissociate the effort associated with successful listening from the cognitive abilities that affect performance on speech tasks.

In addition to WM, scores on LE tasks may be moderated by other general cognitive traits, such as individual differences in processing speed (Desjardins & Doherty, 2014; Picou, Ricketts, & Hornsby, 2013), inhibition (Mishra et al., 2013b), linguistic closure (Mishra et al., 2013a, 2013b), and updating (Mishra et al., 2013b). For example, Zekveld and colleagues (Zekveld & Kramer, 2014; Zekveld et al., 2011) found that individuals with a greater capacity to complete linguistic units (as measured by the Text Reception Threshold [TRT] test) have greater pupil dilation when speech is difficult to understand. These individuals presumably utilize more cognitive resources during speech understanding, which results in increased processing load and greater pupil dilation (in line with the effort and resource hypotheses). In contrast, Mishra et al. (2013b) showed that better performance on the TRT test was associated with better performance on the CSCT, following the predictions of the cognitive efficiency and ELU hypotheses.

Thus, the literature to date is quite mixed on whether and how multiple cognitive abilities affect the effort needed for successful speech understanding. The current study attempts to systematically assess how cognitive abilities affect the effort required to understand speech by including a large cognitive battery that includes a majority of the cognitive tasks that have been used previously in the LE literature. These include the Rspan task (WM; Daneman & Carpenter, 1980), the letter memory (LM) task (updating; Mishra et al., 2013a, 2013b), the Simon task (inhibition; Mishra et al., 2013a, 2013b), the TRT test (linguistic closure; Mishra et al., 2013a, 2013b), and a lexical decision task (LDT; processing speed; Picou et al., 2013).

Personality. To date, no studies have assessed whether and how personality traits—specifically extraversion and sensory processing sensitivity—affect the amount of LE that participants must expend. However, there is reason to expect that LE may be affected by these personality traits. Eysenck (1967, 1990) hypothesized that introverts are chronically more cortically aroused than extraverts due to differences in resting levels of activation of the ascending reticular activating system. These differences may lead extraverts to seek out additional stimulation to reach an optimal level of arousal. Indeed, Geen (1984) demonstrated that when given a choice of noise level, extraverts opted for louder background noise than introverts while completing a paired-associates learning task. At higher noise levels, introverts showed worse performance on the memory task than extraverts. Thus, introverts may be expected to exert greater LE than extraverts, particularly in challenging listening conditions. The current study included the Big Five Inventory–2 (BFI-2; Soto & John, 2017) to measure extraversion.

Another variable that may be expected to influence LE is sensory processing sensitivity (SPS). SPS is characterized by greater sensitivity to subtle (e.g., low intensity) stimuli and heightened emotional reactivity (Aron & Aron, 1997; Aron, Aron, & Jagiellowicz, 2012). Aron and Aron (1997) showed strong correlations among self-reported sensitivities, including sensitivity to caffeine, hunger, and strong sensory input, among others. Individuals with high SPS are believed to have lower perceptual thresholds and process stimuli more deeply than others (see Jagiellowicz et al., 2011, for neuroimaging evidence for greater visual processing in high SPS individuals). As a result, these individuals are likely to be overstimulated by external sensory stimuli. We hypothesized that participants who score higher on SPS may have more difficulty with the LE tasks and may be particularly impaired by an increase in background noise (see Boothroyd & Schauer, 2015, for a demonstration that listeners have different "noise tolerance profiles" that affect their priorities while listening to speech in noise). Indeed, some questions on Aron and Aron's (1997) measure of SPS refer directly to participants' emotional reactions to loud noises or intense sensory input, which seems highly related to those intending to assess participants' responses to increased task difficulty in listening situations. We measured SPS using the standard measure in the literature, the Highly Sensitive Person Scale (HSPS; Aron & Aron, 1997).

Method

Participants

A total of 111 members of the Carleton College community aged 18–28 years with normal hearing and normal or corrected-to-normal vision completed a battery of LE and cognitive tests. Testing took approximately 2.5 hr, spread over two sessions (1.5 hr for the first session, 1 hr for the second session) that were completed an average of 4 days apart. All methods were approved by the Carleton College Institutional Review Board.

Design

In an initial session, participants completed seven LE tasks. In the second session, they completed an audiologic screening, five cognitive tasks, and two personality measures. The LE measures were administered in six orders using a Latin Squares design, and participants were randomly assigned an order. Six rather than seven orders were included because the subjective, self-report measure of LE was collected multiple times during each of the other LE tasks. Although it may be preferable to use a consistent, pseudorandomized order for all tasks in individual differences studies, given that we were also interested in assessing the sensitivity of the LE measures, we opted to counterbalance the tasks to avoid order effects. However, because we were not interested in assessing the sensitivity of the cognitive tasks, they were presented in a consistent order in the second session: the Rspan task (Conway et al., 2005; Daneman & Carpenter, 1980), the Simon task (Mishra et al., 2013a, 2013b; Simon, 1969), the LM task (Morris & Jones, 1990), the LDT (Meyer & Schvaneveldt, 1971), the TRT test (Zekveld, George, Kramer, Goverts, & Houtgast, 2007), the HSPS (Aron & Aron, 1997), and the BFI-2 (Soto & John, 2017). Table 1 shows a brief description of all the tasks.

Stimuli

Stimulus lists, auditory speech materials, and files for stimulus presentation are freely available for reuse at https://osf.io/8z4ef/. All speech stimuli were recorded by the same female speaker with a standard midwestern accent. For the CSCT, an additional male speaker was used. Speech stimuli were edited and matched on root-mean-square amplitude using Adobe Audition (2017).

Although LE may be induced in a number of ways, given the ubiquity of background noise in common listening situations and the ease with which noise may be added to any LE task, we opted to manipulate the amount of LE required by changing the SNR. Some prior studies have individually adapted SNR to equate word recognition performance across participants (e.g., 80% accuracy in Picou & Ricketts, 2014), but others have used a consistent SNR for all participants (e.g., Pichora-Fuller et al., 1995). Given that the goal of the current study was to assess individual differences in LE, we did not individualize SNR out of concern that variable noise levels would complicate interpreting individual differences in LE tasks and comparing them with individual differences in cognitive and personality measures. Thus, we chose fixed SNRs that most closely approximated the range that has been used in recent work (Keidser et al., 2015; Mishra et al., 2013a; Picou & Ricketts, 2014). Based on pilot testing of individual, monosyllabic words from the CDT (using different participants than in the actual experiment and omitting the secondary task), we opted for an SNR of +5 dB for the easy, low-noise listening condition and −2 dB for the hard, high-noise condition. We attempted to choose a hard SNR that was sufficiently difficult to require effort but not so difficult that it caused listeners to give up due to cognitive overload (see Wu et al., 2016; Zekveld & Kramer, 2014). Although using a fixed SNR is best suited for our purposes, it means that intelligibility levels may differ across LE tasks, as a consistent change in SNR will differentially affect difficulty depending on stimulus materials, set size, and other factors

(Kryter, 1970). In all LE tasks, the masker was speech-shaped noise, created using Praat (Version 6.0.36) to match the long-term average spectrum of the female talker, and was presented continuously throughout the task. Following the convention of the majority of work on LE (Desjardins & Doherty, 2014; Fraser et al., 2010; Gosselin & Gagné, 2011a, 2011b; Picou & Ricketts, 2014), the level of the speech stimuli was held constant and the level of masker noise varied to create the different SNRs. The speech was presented at 65 dB SPL, and noise was set to 60 dB SPL in the easy condition and 67 dB SPL in the hard condition. Stimuli were presented binaurally via Seinheisser HD 280 Pro headphones.

A limitation to presenting the easy and hard conditions in a simple blocked design is that participants may improve on the tasks over time. Thus, participants who complete the hard condition second may fail to show effects of increased LE with greater noise because they have gained familiarity with the task. Therefore, we included multiple blocks of easy and hard conditions for each participant and counterbalanced the order of the noise conditions across participants. Using blocked rather than interleaved trials also enabled us to collect self-report measures for each condition (see below).

Session 1 Procedure: LE Battery

Given that the current study is concerned with validating previously used measures of LE, we attempted to follow the methods used in prior studies as closely as possible. Stimulus presentation and data collection were conducted using SuperLab 5 (Cedrus) unless otherwise specified.

Subjective Measure (NASA-TLX)

We used four questions from the original NASA-TLX, omitting two questions related to physical demand and temporal demand. For each question, an unnumbered scale with 21 subdivisions from low on the left to high on the right appeared on the screen, and participants clicked the section of the scale that corresponded to their experience during the previous block or trial (see descriptions of tasks below). The next question then appeared until the participant reported on each of the four questions. The order of these questions was consistent (mental demand, effort, performance, frustration). Self-report scores for the effort question alone and for all four questions averaged were calculated for each condition of each task, following the procedures of Seeman and Sims (2015). Scores for "performance" were reverse coded to put all measures on a consistent scale (higher values indicate more effort/demand/frustration/poorer performance) prior to averaging. NASA-TLX scores that were collected from participants as they completed all the other LE measures (CDT, SDT, RM, LS, CSCT, and pupillometry) were combined prior to analysis.

After the four NASA-TLX questions were presented, participants were also asked, "how often did you give up on trying to perceive the speech?" and responded on a 21-point scale (see Zekveld & Kramer, 2014, for prior work on assessing giving up). Although Zekveld and Kramer (2014) used a 10-point scale, we opted for a 21-point scale to keep all subjective measures consistent and reduce participant confusion. Zekveld and Kramer (2014) suggest that, when listening conditions are too difficult, participants experience "cognitive overload" and give up on the task, so they expend less effort than they did when the task was difficult but not impossible. Although the SNR used in the current study was well above the level that likely results in cognitive overload, including a measure of "giving up" enabled us to exclude participants when they were not fully attending to the listening task.

Dual-Task Measures

For both types of dual-task paradigms, the primary task was listening to words. Participants verbally reported the word they heard and were instructed to prioritize the word recognition task (Bourland-Hicks & Tharpe, 2002; Desjardins & Doherty, 2013; Downs, 1982; Fraser et al., 2010). Word lists were obtained from Picou and Ricketts (2014) and included monosyllabic words (e.g., "that," "base," "low"). In each task, participants were presented with four blocks of 30 words, half in the easy condition and half in the hard condition, with difficulty alternating and the starting difficulty counterbalanced across participants. The word lists were counterbalanced such that half the participants heard a given word in the hard condition and half heard it in the easy condition, and different word lists were used for the CDT and the SDT. NASA-TLX subjective reports were taken after each of the four blocks of 30 words.

CDT. In the CDT, words were presented at variable intervals of 2,000–3,000 ms (in 500 ms intervals). In addition to completing the primary word recognition task, participants also completed a secondary visual task. Two square boxes measuring approximately 5 cm were shown on the screen, and a digit between one and eight appeared in one of the boxes at quasirandom intervals (500–2,000 ms after the end of the previous trial, in 500 ms steps). Following the procedures of Sarampalis et al. (2009), participants responded by pressing either a button with a left-facing arrow or one with a right-facing arrow on a button box (Cedrus RB-740). They pressed the button pointing away from the square with the number for odd numbers (e.g., if the number 7 appeared in the right box, participants pressed the left-facing arrow) and the button pointing toward the square with the number for even numbers. The number remained on the screen until the participant responded or 2,500 ms elapsed. Because the response latencies and interstimulus intervals were variable, the visual stimuli did not systematically coincide with aurally presented words. Participants were instructed to respond as quickly and accurately as possible, and reaction time and accuracy were recorded for each trial. Although Picou and Ricketts (2014) included twice as many word trials in their CDT as the current study, their secondary task included many fewer of the critical probe trials (15) that contribute to the dual-task LE measure. Sarampalis et al.

(2009) presented approximately 200 s of audio with the same spacing of visual probes used in the current study, so the number of stimuli is comparable to the number we used. Prior to starting the task, participants completed 10 s of practice on the secondary odd/even classification task alone, followed by practice on the word recognition with the secondary task (five words, approximately 15 s). All practice trials were completed with the experimenter in the room, and practice was repeated once if participants appeared not to understand the task after their first attempt at the practice trials. CDT scores for individual participants were calculated as the average reaction time to visual stimuli for correct responses. Although some studies have collected data on primary and secondary task performance in isolation (Gosselin & Gagné, 2011a, 2011b), we opted not to because of difficulties using those data to compare tasks. For instance, although we could have collected data on secondary task performance in isolation for the CDT, no analogue exists for the SDT (see below).

SDT. The SDT paradigm required participants to make judgments about the words they heard in the primary task. After each word, they were asked to decide as quickly and accurately as possible whether the word was a noun and indicate their response via button press. After doing so, they verbally repeated the word. Words were presented with 3,000 ms of silence between them. Noun categorization reaction time was recorded for each trial. Participants completed 10 practice trials prior to beginning. Following the procedures of Picou and Ricketts (2014), SDT scores for individual participants were calculated as the average reaction time for all noun judgment responses (not just correct responses) because most nouns can also be categorized as other parts of speech (approximately 84% according to Picou & Ricketts, 2014), and participants may differ in the ability with which they can categorize nouns (i.e., semantic judgment accuracy was not scored).

Recall Measures

RM. The RM task was based on the methods of McCoy et al. (2005; see also Brault, Gilbert, Lansing, McCarley, & Kramer, 2010; Bunting, Cowan, & Saults, 2006; Sommers & Phelps, 2016; Sommers, Tye-Murray, Barcroft, & Spehar, 2015). Participants heard 16 lists that included five, seven, eight, 10, 12, 13, 14, or 15 words (two of each length). Words were presented with 1,000 ms of silence between them, and list order was pseudorandomized (so participants were unaware of the list length on each trial; Sommers & Phelps, 2016). These lists were divided into four blocks, split between easy and hard conditions, with difficulty alternating and the starting difficulty counterbalanced across participants. The word lists were counterbalanced such that half the participants heard a given list in the easy condition and half heard it in the hard condition. NASA-TLX subjective reports were taken after each of the four blocks. Blocks and list positions were matched on lexical variables, including word frequency and neighborhood density. Words were taken from the set used by McCoy et al. (2005), and the order was randomized to avoid semantic relatedness. After hearing all words in a list, participants saw three asterisks appear on the screen (to avoid exposure to additional verbal material; McCoy et al., 2005; also see Brault et al., 2010), which indicated that they should verbally repeat the last three words they heard in any order. Given that participants were unaware of the length of each list, they had to continuously update the last three words in memory. Participants did not repeat the words aloud as they heard them.

McCoy et al. (2005) and Sommers and Phelps (2016) presented the RM task in the absence of background noise and calculated RM scores as the percent of two- and three-back words correctly perceived. In the absence of noise, intelligibility was assumed to be high (and ceiling level performance on 1-back scores support this), so group differences in two- and three-back performance likely represent differences in encoding rather than intelligibility. Given that the task was presented in noise in the current study, it is necessary to account for the overall audibility of the words in the easy and hard noise conditions to ensure that poorer observed performance in the hard condition is a function of poorer encoding, not simply poorer intelligibility. Therefore, adjusted scores were calculated in which scores on the two- and three-back words were divided by the average accuracy on the 1 back word in each noise condition for each participant. This indicates the extent to which performance on two- and three-back words is worse than would be expected on the basis of the intelligibility of the words. Thus, higher numbers still indicate better performance, with scores of 1.0 indicating that two- and three-back words were recalled at equivalent accuracies as 1-back words.

LS. Participants were presented with high-predictability (HP; e.g., "the watchdog gave a warning growl") and low-predictabililty (LP; e.g., "the old man discussed the dive") sentences from the Revised Speech Perception in Noise lists (Bilger, Nuetzel, Rabinowitz, & Rzeczkowski, 1984; Kalikow, Stevens, & Elliott, 1977). Following the presentation of each sentence, participants made a judgment via key-press indicating whether or not the word was predictable from the preceding context (to ensure that the entire sentence was processed rather than only the final word; Daneman & Carpenter, 1980; Pichora-Fuller et al., 1995), then verbally reported the final word. The amount of time allotted to respond to each sentence was limited to 2,000 ms after sentence offset to minimize opportunities for strategic rehearsal. There was a 1,000 ms interstimulus interval between the button press and the onset of the next sentence. After lists of four, six, or eight sentences, participants were prompted to recall all final words in the set in any order (lists of two sentences or fewer were not included because young adults typically have little difficulty recalling both final words in a two-item set; Pichora-Fuller et al., 1995). Words were scored as correct if they were recalled as they were perceived; that is, a participant could incorrectly identify a word and still receive credit for correctly recalling it if they recalled what they initially misperceived (Johnson et al., 2015; Ng et al., 2013; Pichora-Fuller et al., 1995; Sarampalis et al., 2009).

Participants heard a total of 144 sentences that were split into lists of four, six, or eight sentences (eight of each length, for a total of 24 lists), with half in each noise condition. There were six blocks such that noise condition changed every four lists, with half the participants starting with easy and half starting with hard. Lists were counterbalanced such that each list was presented in the hard condition for half the participants and in the easy condition for the other half. NASA-TLX questions were presented after each of the six blocks. Following the procedures of Pichora-Fuller et al. (1995), lists progressed from shorter to longer throughout the task. However, given that we were interested in analyzing the effect of noise level on recall performance, we counterbalanced the noise condition across participants. This is a slight deviation from Pichora-Fuller et al. (1995) and Daneman and Carpenter (1980), in which all lists were presented in ascending order in the easier noise condition, then all lists were presented in the more difficult noise condition. This change was necessary to enable us to evaluate the effect of noise on LE without confounding list length and noise condition and to counteract possible order effects.

The order in which sentences were presented in each list was pseudorandomized, HP and LP sentences were intermixed (with at least one HP and one LP sentence per list), and the number of HP and LP sentences was matched across the noise conditions. Following the procedures of Pichora-Fuller et al. (1995), list size varied (Daneman & Carpenter, 1980; Zekveld & Kramer, 2014) out of concern that keeping a consistent set size (as in Johnson et al., 2015; Ng et al., 2013; Sarampalis et al., 2009) could lead to strategic biases driven by participants' knowledge of how many words they can easily keep in their WM. Participants completed two practice lists (of lengths four and six). Individual LS scores were calculated as the proportion of words correctly recalled (or recalled as perceived).

CSCT. The CSCT was based on the procedures used by Mishra et al. (2013a, 2013b). Participants heard numbers between 13 and 99 with 1,000 ms of silence between them spoken by a male and a female voice. They were asked to monitor the highest or lowest number spoken by each talker (updating condition) or to recall the odd or even numbers spoken by one talker (inhibition condition) in low-load (recalling two items) and high-load (recalling three items, including the first number) conditions (see Keidser et al., 2015, for variations on this task). Participants heard 32 lists total, divided evenly between the noise (easy and hard), executive task (updating and inhibiting), and load (low-load and high-load) conditions, resulting in four lists in each of the eight conditions (see Mishra et al., 2013b, for a helpful schematic of the CSCT task). Executive condition was blocked and counterbalanced across participants, and load varied on each trial. Participants received instructions before each list about how many and which numbers to attend to. Within each executive task, the noise and load conditions were pseudorandomized. After the 13 numbers in each list were presented, participants verbally reported the two or three numbers they had been instructed to recall.

NASA-TLX measures of LE were presented after eight of the trials (that were selected to include two of each type of executive task, noise, and load condition). Note that this differs from presentation of the subjective measures on other tasks, in which participants responded to the difficulty of a whole block rather than an individual trial. This change was included to allow for pseudorandomization of trial order (in line with prior work), and participants were explicitly made aware of the change. CSCT scores were calculated as the proportion of numbers correctly recalled in each of the conditions.

Physiological Measure (Pupillometry)

Participants were seated approximately 55 cm away from a 61-cm PC monitor. Stimulus presentation and data collection were controlled via Tobii Studio Professional (Version 3.2; 2013). Average pupil size of both eyes was measured using a Tobii-X260 Eye Tracker, wide edition. The room was moderately lit (approximately 70 fL), and the background of the screen throughout the task was grey ([125, 125, 125] in red/green/blue space) to generate intermediate pupil size. Maximum and minimum pupil sizes were calculated by presenting plain black and white screens for 10 s each at the beginning of the task (Piquado et al., 2010). These values were used to confirm that pupil responses on the test trials were above floor and below ceiling levels (see Results). Calibration was done with Tobii's built-in 9-point calibration system. Following the procedures of Winn et al. (2015), audio stimuli consisted of 40 sentences from the Institute of Electrical and Electronics Engineers/Harvard sentence corpus (Rothauser et al., 1969). Participants heard four blocks of 10 sentences with five key words each, half in the easy condition and half in the hard condition, with difficulty alternating and the starting difficulty counterbalanced across participants. The sentence lists were counterbalanced such that half the participants heard a given sentence in the hard condition and half heard it in the easy condition. NASA-TLX subjective reports were taken after each of the four blocks. Participants completed five practice trials prior to beginning the task.

Background noise was presented continuously throughout each block, and participants were instructed to fixate on a black cross at the center of the screen. The sentence began after a 3,000 ms delay (Zekveld & Kramer, 2014) and lasted an average of 2,811 ms. Two thousand milliseconds after speech offset, a 1,000 ms long 1000 Hz tone prompted participants to repeat back the sentence they just heard. Participant responses were recorded and later coded for the number of keywords (out of five) that were correctly identified (following Winn et al., 2015). Due to a recording error, one sentence contained only four keywords. Participants initiated the next trial by clicking the mouse.

For each trial and each participant, we determined the peak pupil size in the interval between sentence onset and the response prompt (Koelewijn et al., 2015; Koelewijn, Zekveld, Festen, & Kramer, 2012; Zekveld et al., 2014). Average baseline pupil size (collected during the 1,000 ms interval prior to sentence onset; Koelewijn et al., 2015; Zekveld et al., 2014; Zekveld & Kramer, 2014) for each trial was subtracted from the trial peak to arrive at peak pupil dilation. Because the critical value for the window of interest was the peak value, we did not attempt to remove frames with blinks (which may result in inaccurately small values).

Session 2 Procedure: Audiologic, Cognitive, and Personality Battery

Audiogram

Pure-tone thresholds (PTA) were determined for both the left and right ears at 250, 500, 1000, 2000, 4000, and 8000 Hz using a Maico MA-39 audiometer. All tests were conducted in a sound-attenuating chamber.

Rspan Task (WM Capacity)

Participants completed a standard Rspan task (Conway et al., 2005; Daneman & Carpenter, 1980; Foo, Rudner, Rönnberg, & Lunner, 2007; see Lunner, 2003; Mishra et al., 2013a, 2013b; Rudner et al., 2012; Rudner, Rönnberg, & Lunner, 2011; for demonstrations of Rspan in the LE literature). Participants were visually presented with declarative sentences that were sensical (e.g., "During winter you can get a room at the beach for a very low rate") or nonsensical (e.g., "During the week of final spaghetti, I felt like I was losing my mind") on a computer screen. Stimuli were obtained from the Engle lab (Redick et al., 2012). Participants silently read each sentence and judged whether it was semantically sensical or absurd by pressing a button (in order to ensure that participants read each sentence and not just the final word; Mishra et al., 2013a, 2013b). After a 1,750 ms interstimulus interval, another sentence was presented. After three, four, five, or six sentences, participants were instructed to recall verbally the final word of each sentence. They then initiated the next set via a button press. Three trials of each span length (three, four, five, and six sentences) were presented in an ascending order, and span was calculated by adding together the number of words correctly recalled out of each set, giving a max score of 54, which was converted to proportion correct prior to analysis. Participants completed three practice trials (lengths three, four, and five) prior to beginning.

LM Task (Updating)

This task was based on the LM task used by Mishra et al. (2013a, 2013b; adapted from Miyake et al., 2000; Morris & Jones, 1990). Participants saw lists of five, seven, nine, or 11 consonants one at a time on the center of the screen for 2 s each and were asked to hold the last four in memory. At the end of each sequence, they were prompted to recall the last four in any order and indicate their responses using the keyboard. Because participants did not know the length of each list, they were required to continually update the last four letters in memory. Participants practiced on two lists of lengths seven and nine (Mishra et al., 2013a). Testing consisted of 12 randomized lists (three of each length). The score was the proportion of letters correctly recalled, irrespective of order (Mishra et al., 2013a, 2013b).

LDT (Processing Speed)

Processing speed was measured using a standard LDT, based on the one used by Picou et al. (2013). Participants were presented with a four-letter orthographic string (e.g., "SHIP" or "SIRT") and were asked to determine as quickly and accurately as possible whether it formed a real English word and indicate their response via button press on a buttonbox. Words were common (log frequencies greater than 3 according to the SUBTLEX-US corpus—a database containing word frequency information collected from 51 million word tokens from film and television program subtitles; Brysbaert, New, & Keuleers, 2012), and nonwords were derived from single-letter substitutions of other equally common words. The interstimulus interval was 1,000 ms. Participants completed 100 trials (half words, half nonwords), presented in a randomized order, preceded by five practice trials. Processing speed was quantified as average reaction time for correct trials.

TRT Test (Linguistic Closure)

The TRT test (Besser et al., 2013; Besser, Zekveld, Kramer, Rönnberg, & Festen, 2012; Mishra et al., 2013a, 2013b; Zekveld et al., 2007) is often taken as a visual analogue to the speech reception threshold test (Zekveld et al., 2007), as both tests require that participants form perceptual wholes from incomplete auditory or visual information. In the TRT test, participants were presented with sentences from the Institute of Electrical and Electronics Engineers/Harvard sentence corpus (Rothauser et al., 1969; note that we selected different sentences for the TRT than were used for the pupillometry test) obscured to varying degrees by vertical black bars. First, the black bars appeared on a white background, and then each sentence was presented one word at a time in red font at approximately the rate it would take to produce the sentences, and participants had 3,500 ms to read the partially masked sentence aloud (Zekveld et al., 2007). An adaptive one-up, one-down procedure was used in order to determine the percentage of visible text required for a participant to identify 50% of all sentences completely correctly (Mishra et al., 2013a). The first sentence was presented with a visibility of 40%, and masking was decreased by 12% until the first sentence was identified completely correctly. After the initial sentence was identified correctly, sentences 2–13 were presented only once with a step size of 6% (following the methods of Mishra et al., 2013b; Zekveld et al., 2007). Threshold was calculated by determining the average percentage of unmasked text for all sentences other than the first sentence (including what would be presented for the 14th sentence). Thus, low TRT score indicates better performance.

Simon Task (Inhibition)

This task was based on the Simon task used in Mishra et al. (2013a, 2013b; see also Pratte, Rouder, Morey, & Feng, 2010). On each trial, participants were presented with a red or blue rectangle on the right or left side of the screen, presented at 1,000 ms intervals. They were asked to respond as quickly as possible to red blocks by pressing a red key on the right side of the button box and to blue blocks by pressing a blue key on the left side (ignoring the spatial location of the block itself). On congruent trials, the spatial location was the same as the correct response key (i.e., red on the right and blue on the left). Inhibition was measured as the difference in average reaction time for correct trials between incongruent and congruent trials, such that higher values indicate a greater incongruency cost, and thus, poorer inhibition.

BFI-2 (Extraversion)

The BFI-2 is a standardized self-report measure that includes 60 questions with the carrier phrase "I am someone who…" (e.g., "is outgoing, sociable" or "is compassionate, has a soft heart"; Soto & John, 2017) to assess five domains of personality (including extraversion). Participants were presented with individual questions on a computer screen and responded using a 5-point scale (1 = *disagree strongly,* 2 = *disagree a little,* 3 = *neutral*; *no opinion,* 4 = *agree a little,* 5 = *agree strongly*). Given our hypothesis that extraversion may be related to LE, average composite extraversion scores were calculated, but scores for the other measures of personality were not analyzed. However, we opted to administer the full test to help avoid alerting participants to the purpose of the task.

HSPS (Sensitivity)

The HSPS is a standardized self-report measure that consists of 27 questions (see Aron & Aron, 1997, for the full list of questions). The HSPS procedure was the same as that for the BFI-2, except the scale was 1–7 with anchors *not at all* (1), *moderately* (4), and *extremely* (7). Total HSPS scores were calculated by averaging all responses.

Results

Tasks that required participants to give verbal responses were coded offline by experimenters. In cases where a particular response was ambiguous, a second experimenter was consulted. Analyses were conducted using R (Version 3.4.0; R Core Team, 2016). All raw data and scripts to run analyses are available at https://osf.io/8z4ef/.

Data Loss and Cleaning

Two participants did not return for the second session. Their data for the first session were used, resulting in data from 111 participants for the LE measures and 109 for the cognitive, personality, and audiologic measures. Data for some tasks for some participants were lost due to experimental error (e.g., failing to record audio stimuli, poor calibration of the eye tracker) or participants' failure to follow instructions. These accounted for approximately 1% of all tasks.

An additional two participants' data were excluded from the CDT because of low levels of accuracy (more than 3 *SD*s below the mean accuracy) on the secondary task. On average, participants responded correctly to the odd/even numbers in the CDT at high rates (*M* = 94%, *SD* = 6%). Participants' data for the LS and Rspan tasks were also excluded if they had low accuracies (more than 3 *SD*s below the mean accuracy) for the predictability and sensicality judgments. Poor performance on these tasks may indicate that the participant was not attending to the full sentence and instead was only trying to identify the final word. Accuracies on these tasks were generally high (LS: *M* = 87%, *SD* = 10%; Rspan: *M* = 88%, *SD* = 9%), and only one participant for the LS task and two for the Rspan task were removed. In addition, for all tasks that involved reaction time (CDT, SDT, and LDT), individual reaction times more than 3 *SD*s from each participant's mean for that task were excluded.

As part of each subjective measure block, participants were asked how frequently they gave up on the listening task (Zekveld & Kramer, 2014). Given that measures of LE are designed to assess the cognitive costs of listening, blocks in which participants decide to abandon the listening task do not accurately represent the difficulty of listening. For example, if a participant in a difficult listening situation opts to divert resources from the listening task in the CDT, they may perform better on the secondary task than if they were attending to the listening task. Participants in our study rarely reported abandoning the task—the average "give up" score out of a maximum of 21 was 3.63 (*SD* = 3.59). Given the very low mean and standard deviation, we set a more conservative criterion for removing data than 3 *SD*s above the mean. Tasks on which a participant reported giving up using the top quartile of the scale (54 tasks or 7% of all remaining LE tasks) were removed from analysis.

To assess our attempt to generate intermediate pupil size for the pupillometry task, we compared each participant's average pupil size during the critical window (the point in the task that we expected pupil size to be largest) to the maximum size generated when presented with a blank, black screen. All participants had smaller pupils during the critical window than when presented with a black screen. In addition, we calculated the average size during the baseline (the point in the task that we expected the smallest pupil size) and the minimum size when presented with a blank white screen. All but two participants had larger pupils on average during baseline than the minimum when presented with a white screen. These results confirm that pupil size was not systematically limited by the floor or ceiling.

Descriptive Data

Descriptive data for the LE tasks in the easy and hard conditions are shown in Table 3. Variables were checked for normality using the Shapiro–Wilk test and, as needed, transformed to help them more closely approximate a normal distribution before calculating *t* tests. Given the large number of comparisons, we used the Holm–Bonferroni correction using the R function p.adjust (*stats* package) when calculating *p* values.

Table 3. Descriptive statistics for listening effort measures and *t* tests comparing the easy (SNR = +5 dB) and hard (SNR = −2 dB) conditions for each task.

		Easy condition		Hard condition		Test statistics		
Task	*N*	*M* (*SD*)	Range	*M* (*SD*)	Range	*t* value	*p* value	Cohen's *d*
NASA-TLX	111	10.63 (2.24)	5.00–15.3	12.39 (2.31)	5.57–18.02	−16.00	**< .001**	0.77
NASA-TLX, effort	111	12.46 (2.92)	5.15–19.86	13.90 (2.8)	6.18–19.53	−11.25	**< .001**	0.50
CDT	103	802 (158)	512–1,392	834 (177)	520–1,440	−5.44	**< .001**	0.18
SDT	105	1,249 (187)	849–1,687	1,310 (201)	898–1,882	−5.93	**< .001**	0.31
CSCT	90	0.84 (0.11)	0.53–1.00	0.84 (0.12)	0.59–1.00	−0.09	> .99	0.01
CSCT, updating	90	0.79 (0.14)	0.44–1.00	0.79 (0.15)	0.44–1.00	−0.29	> .99	0.03
CSCT, inhibiting	90	0.90 (0.12)	0.31–1.00	0.88 (0.13)	0.31–1.00	0.68	> .99	0.07
CSCT, low load	90	0.92 (0.09)	0.56–1.00	0.90 (0.10)	0.62–1.00	1.89	.31	0.24
CSCT, high load	90	0.76 (0.16)	0.31–1.00	0.77 (0.17)	0.38–1.00	−0.84	> .99	0.03
LS	97	0.53 (0.10)	0.32–0.74	0.50 (0.10)	0.31–0.78	5.29	**< .001**	0.33
LS, LP	97	0.45 (0.12)	0.22–0.75	0.40 (0.12)	0.17–0.72	4.81	**< .001**	0.34
LS, HP	97	0.62 (0.10)	0.42–0.83	0.60 (0.11)	0.36–0.83	2.78	**.039**	0.24
RM	98	0.89 (0.11)	0.50–1.14	0.82 (0.19)	0.43–1.62	4.09	**.001**	0.54
Pupillometry	104	0.41 (0.20)	0.11–1.60	0.45 (0.19)	0.10–1.41	−3.59	**.003**	0.27
CDT, word	102	0.91 (0.04)	0.83–1.00	0.72 (0.06)	0.52–0.82	29.92	**< .001**	3.66
SDT, word	103	0.92 (0.04)	0.78–1.00	0.70 (0.08)	0.42–0.83	28.98	**< .001**	3.46
Pupillometry, word	101	0.96 (0.06)	0.68–1.00	0.81 (0.11)	0.54–1.00	14.07	**< .001**	2.00

Note. The top four panels contain descriptive statistics for the listening effort measures (subjective measures, behavioral reaction time measures, behavioral recall measures, and the physiological measure; refer to Table 1 for a summary of each task). Tasks that produce independent measures for word recognition are shown in the bottom panel. Values represent proportion of words correct, with the following exceptions: NASA-TLX ratings are scores out of 21, CDT and SDT values are reaction times in ms, and pupillometry values are peak pupil dilation in mm. *p* values that were not significant before the Holm–Bonferroni correction are not returned, so they are indicated here as *p* > .99. Cohen's *d*s are given in absolute values. SNR = signal-to-noise ratio; NASA-TLX = National Aeronautics and Space Administration Task Load Index; CDT = complex dual-task; SDT = semantic dual-task; CSCT = Cognitive Spare Capacity Test; LS = listening span; LP = low predictability; HP = high predictability; RM = running memory.

Note that sample sizes differ across tasks, primarily due to differences in the rates at which participants reported giving up on the listening task. As expected, performance was typically better (e.g., faster reaction times, higher accuracy), and pupil size was smaller in the easy condition than in the hard condition. The CSCT was the only LE task that did not show significant differences between the easy and hard conditions. Previous research has found that CSC is sensitive to changes in noise (Keidser et al., 2015; Mishra et al., 2013a, 2014), but these studies compared performance in quiet to that in various types of noise at individualized SNRs. Thus, the current study is the first to demonstrate that the CSCT appears not to be sensitive to changes in SNR, at least in steady-state, speech-shaped noise at the SNRs used here.

Descriptive data for the cognitive tasks are shown in Table 4. All measures showed good variability and typically had values that are comparable with what has been reported previously for these tasks in the LE literature (Benham, 2006; Mishra et al., 2013a, 2013b, 2014; Soto & John, 2017).

Aim 1: Convergent Validity

We first calculated average scores for each participant for each task, collapsing across SNR, in order to obtain a general metric for how successfully the participant completed the task. We then calculated correlations among those values, correcting for multiple comparisons using the rcorr.adjust function (*RcmdrMisc* package; Holm–Bonferroni method). Variables were checked for normality and transformed as described above. In line with prior work (Mishra et al., 2013a, 2013b, 2014), scores for the CSCT were rationalized arcsine transformed. Correlations

Table 4. Descriptive values for audiologic, cognitive, and personality measures.

Task	*N*	*M* (*SD*)	Range
PTA	109	3.85 (4.22)	−5.00–23.33
Rspan	102	0.69 (0.17)	0.22–0.96
LM	109	0.86 (0.08)	0.50–1.00
LDT	109	606 (114)	452–1425
LDT, acc	109	0.97 (0.03)	0.86–1.00
Simon	109	18 (28)	−112–77
Simon, acc	109	0.95 (0.05)	0.74–1.00
TRT	109	66.34 (4.09)	58.00–74.62
Extraversion	107	3.28 (0.79)	1.67–5.00
HSPS	108	4.53 (0.72)	3.00–6.59

Note. PTA = average threshold for the better ear; Rspan = proportion of words correctly recalled on the reading span task; LM = proportion of letters correctly recalled on the letter memory task; LDT = reaction time in ms to correct trials on the lexical decision task; LDT, acc = accuracy on the LDT task; Simon = incongruency cost in ms on the Simon task; Simon, acc = accuracy on the Simon task; TRT = score on the Text Reception Threshold test; Extraversion = mean extraversion rating out of 5; HSPS = mean sensitivity rating out of 7 on the Highly Sensitive Person Scale.

among the LE measures are shown in Table 5, below the diagonal.

We report all the subcomponents of the CSCT, the NASA-TLX, and the LS task here to follow the convention of prior studies (e.g., Mishra et al., 2013a). Unsurprisingly, the strongest correlations were between these different measures from the same task. For example, the correlations between the CSCT updating and low-load conditions are artificially inflated because some of the same trials are included in each measure. The reaction time measures (the CDT and the SDT) were significantly correlated, and most components of the recall measures (the CSCT and the LS task) were significantly correlated as well (with the nonsignificant correlations in the expected direction and nearing significance). The other recall measure, RM, was not significantly correlated with any component of the CSCT or LS task, or with the reaction time–based measures, but was negatively correlated with the effort portion of the NASA-TLX measure; participants who performed well on the RM task tended to report less subjective effort overall. Peak pupil dilation was the only measure to not show any significant correlations with other indices of LE.

The previous correlational analysis tested whether participants' scores on LE measures correlated (collapsing across difficulty conditions). A related question is whether the detrimental effect of noise is relatively consistent across tasks. That is, are participants who tend to be strongly affected by noise in one task also strongly affected by noise in another task? To assess this, we built a linear mixed-effects model (via the *lmer4* package in R, Version 3.4.0) to predict task scores. Dependent variables that were proportions (CSCT, LS, RM) were logit transformed, and then all variables were transformed to z-scores so that they could be entered into the model simultaneously. These scaled scores from all tasks were then predicted using task, difficulty (easy vs. hard), and the critical Task × Difficulty interaction as fixed effects and participants and items as random effects. We first compared this full model to a reduced model that excluded the critical Task × Difficulty interaction using a likelihood ratio test. This test indicated that the full model fits the data better than the reduced model ($\chi^2_{11} = 214.04$, $p < .001$), suggesting that there were differences in how the noise manipulation affected tasks. We then reran the full model multiple times, once with each task as the reference category, to enable every possible pairwise comparison and therefore determine which task pairs were differently affected by noise. Absolute values of t values generated by the models for each pair of tasks are shown above the diagonal in Table 5. Note that significant values indicate that the effect of noise condition (easy vs. hard) differed for a pair of tasks.

The results suggest that noise did not have uniform effects across tasks. That is, participants who were more affected by noise on one task were not necessarily more affected by noise on others. The exceptions to this were some recall measures; for instance, the effect of noise was less inconsistent across the CSCT and LS task, and the RM task to some extent. In addition, the effect of noise was not significantly different for the pupillometry task and the reaction time measures, CDT and SDT.

Aim 2: Sensitivity

Measures of effect size (Cohen's *d*) were calculated for each task, collapsing across participants (see Johnson

Table 5. Correlation matrix showing relationships among LE measures (collapsing across easy and hard conditions; lower panel) and absolute values for *t* tests assessing the interactions between task pairs and difficulty condition (easy vs. hard; upper panel).

Task	NASA-TLX	NASA-TLX, effort	CDT	SDT	CSCT, updating	CSCT, inhibiting	CSCT, low load	CSCT, high load	LS, LP	LS, HP	RM	Pupil
NASA-TLX	—	*1.92*†	7.14^{***}	5.93^{***}	6.89^{***}	7.28^{***}	6.51^{***}	7.99^{***}	9.50^{***}	8.39^{***}	10.12^{***}	5.43^{***}
NASA-TLX, effort	*0.79****	—	4.52^{***}	3.41^{***}	4.93^{***}	5.32^{***}	6.03^{***}	4.55^{***}	7.33^{***}	6.22^{***}	8.12^{***}	3.08^{**}
CDT	0.16	0.07	—	2.10^{*}	2.15^{*}	2.69^{**}	3.69^{***}	1.62	5.62^{***}	3.89^{***}	6.63^{***}	$1.76^{\dagger}$
SDT	0.05	–0.01	0.34^{*}	—	3.04^{**}	3.56^{***}	4.51^{***}	2.53^{*}	6.46^{***}	4.84^{***}	7.35^{***}	0.16
CSCT, updating	0.09	0.09	0.19	0.22	—	*0.40*	*1.13*	*0.39*	$1.93^{\dagger}$	0.79	3.17^{**}	2.92^{**}
CSCT, inhibiting	0.27	0.20	0.39^{*}	0.31	*0.54****	—	*0.73*	*0.78*	1.49	0.35	2.76^{**}	3.40^{***}
CSCT, low load	0.11	0.02	0.38^{*}	0.24	*0.65****	*0.71****	—	*1.52*	0.68	0.46	2.02^{*}	4.28^{***}
CSCT, high load	0.19	0.18	0.26	0.27	*0.89****	*0.77****	*0.56****	—	2.37^{*}	1.23	3.56^{***}	2.45^{*}
LS, LP	0.23	0.21	0.29	0.31	0.52^{***}	0.34	0.35	0.51^{***}	—	*1.30*	1.55	5.92^{***}
LS, HP	0.14	0.12	0.21	0.41^{**}	0.53^{***}	0.39^{*}	0.44^{**}	0.50^{***}	*0.75****	—	2.72^{**}	4.46^{***}
RM	0.30	0.33^{*}	–0.08	–0.08	0.15	0.21	–0.01	0.26	0.28	0.16	—	-6.88^{***}
Pupil	–0.04	0.05	0.06	–0.08	–0.27	–0.04	–0.18	–0.18	–0.08	–0.14	–0.02	—

Note. To reduce the number of comparisons, only the subcomponents (and not the composites) of the CSCT and the LS task are shown here. For ease of interpretation, variables in the lower panel have been transformed here to make the signs across tasks consistent. Thus, positive correlations indicate greater convergent validity; more effort in one task is associated with more effort in another task. Significant values in the upper panel indicate that the noise manipulation affected a task pair differently. Italicized items are subcomponents of the same task. LE = listening effort; NASA-TLX = National Aeronautics and Space Administration Task Load Index; CDT = complex dual-task; SDT = semantic dual-task; CSCT = Cognitive Spare Capacity Test; LS = listening span; LP = low predictability; HP = high predictability; RM = running memory.

$^{\dagger}p < .11$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

et al., 2015, for another example of using Cohen's d in the LE literature). We used the following equation: $\frac{M_1 - M_2}{SD_P}$, where M_1 and M_2 denote the means of the easy and hard conditions, respectively, and SD_P denotes the pooled standard deviation given by the following: $\sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$. Here, n_1 and n_2 represent the sample sizes for the easy and hard conditions, and SD_1 and SD_2 are the standard deviations of the easy and hard conditions. Absolute values for Cohen's d are shown in the rightmost column of Table 3. Note that effect sizes are dependent on the SNRs chosen, so the values reported here should not be compared with those from previous work. That is, we could have generated larger effect sizes simply by making the difference in SNR between the easy and hard conditions larger. Thus, measures of Cohen's d are relative and should only be used to compare the measures to one another and not to other effects in the literature.

All significant effect sizes were in the expected direction, indicating that increasing SNR increases LE. The effect size for the LP portion of the LS task was numerically larger than that for the HP portion, which may reflect reduced LE as a result of semantic constraint. That is, if semantic context reduces LE, then differences in LE between easy and hard SNR conditions are expected to be smaller in HP contexts compared with LP contexts. The SDT was numerically more sensitive to changes in SNR than the CDT, which is consistent with the finding that secondary tasks that require greater depth of processing are more likely to detect changes in LE across noise conditions than those that require shallow processing (see Picou & Ricketts, 2014).

Aim 3: Links With Individual Differences

Some prior research has attempted to assess how cognitive measures relate to LE by correlating performance on individual cognitive tasks with scores on individual LE tasks (Mishra et al., 2013a, 2013b). This approach can help distinguish between hypotheses that predict that superior cognitive abilities are associated with increased effort (effort and resource hypotheses) or decreased effort (cognitive efficiency and ELU hypotheses). In contrast, other work has assessed whether the relationship between cognitive abilities and LE changes as a function of task difficulty (Picou et al., 2013). This approach can help assess whether cognitive abilities have a constant effect on task performance (as predicted by the effort and cognitive efficiency hypotheses) or an effect that is moderated by task difficulty (as predicted by the resource and ELU hypotheses). These two approaches provide different information about the mechanisms underlying LE. The first assesses the extent to which individual tasks draw on similar pools of resources as other, well-established cognitive traits. The second shows how these cognitive traits may protect listeners from the detrimental effects of difficult listening situations. The current study uses both approaches: a correlation analysis for the first approach (see Table 6) and a series of mixed-effects models for the second (see Table 7).

This study is the first demonstration of the correlation between SPS and subjective ratings of effort; more sensitive people tend to report higher subjective difficulty completing the LE tasks. The correlation between performance on the LDT and SDT has not been reported previously and may simply reflect individual differences in processing speed. The relationship between LDT and CDT scores failed to reach significance, but it was in the same direction as that between LDT and SDT. Inhibition on the Simon task was not correlated with scores on any of the LE measures. Consistent with this finding, correlations between Simon and CSCT performance have only reached significance in rare instances (see Mishra et al., 2013a, 2013b, 2014).

The correlations between LS and Rspan scores and between LS and TRT scores have been reported previously (Besser et al., 2013). Given that Mishra et al. (2013a) and Besser et al. (2012) also found that Rspan and TRT are correlated, our results corroborate the relatively robust association between LS, Rspan, and TRT scores. Performance on the CSCT tended to be associated with Rspan performance, a finding that typically fails to reach significance, but the positive correlations between CSCT and LM performance are consistent with prior research (Mishra et al., 2013a, 2013b, 2014).

Correlations among the cognitive and personality measures are not shown, but only two comparisons reached significance: Rspan score was positively associated with LM score ($r = .53$, $p < .001$) and negatively correlated with TRT ($r = -.41$, $p < .001$), indicating that larger Rspans are associated with better LM and TRT scores; both of these correlations have been reported previously (Besser et al., 2012; Mishra et al., 2013a, 2014). The correlation between sensitivity and extraversion was numerically stronger than has been reported previously (Smolewska, McCabe, & Woody, 2006) but did not reach significance ($r = -.33$, $p = .09$).

To assess the simultaneous, unique contributions of each of the individual difference variables, we also created models to predict each measure of LE using all the cognitive, personality, and audiologic measures. We used the *lme4* package and checked for multicollinearity among the predictor variables (Frank, 2014), and found that all variance inflation factors were less than 2, which is well below the commonly used benchmarks to determine whether multicollinearity is high (Clark-Carter, 2009). We used treatment contrasts, in which the easy condition was coded as 0 and the hard condition was coded as 1, and a separate model was created for each of the LE tasks. Missing values for cognitive measures ($N = 10$) were imputed from the other cognitive and personality predictors using Multivariate Imputation by Chained Equations (R package *MICE*, Version 2.30). Participants were entered into the model as random effects, and the fixed effects included noise condition (easy or hard), each of the individual difference (cognitive, personality, and audiologic) variables, and the interactions

Table 6. Correlations between average scores on the listening effort tasks (collapsing across easy and hard conditions) and audiologic, cognitive, and personality measures.

Task	PTA	Rspan	LM	LDT	Simon	TRT	Ext.	HSPS
NASA-TLX	−.12	−.17	−.17	.02	.11	.13	−.12	.39**
NASA-TLX, effort	−.22	−.15	−.11	.02	.13	.16	−.14	.27
CDT	.02	.00	−.15	.29	−.14	.07	−.05	.35
SDT	.13	−.24	−.31	.43**	−.17	.27	.05	.03
CSCT, updating	−.19	.48**	.47***	−.21	.03	−.19	.02	−.12
CSCT, inhibiting	−.04	.36	.41*	−.17	−.03	−.14	.01	−.08
CSCT, low	−.10	.39*	.32	−.27	.04	−.06	.16	−.15
CSCT, high	−.12	.47**	.51***	−.15	−.02	−.22	−.05	−.07
LS, LP	.01	.60***	.43**	−.13	−.04	−.40**	.23	−.17
LS, HP	−.10	.57***	.45***	−.17	.04	−.42**	.08	−.03
RM	−.01	.39*	.29	.08	−.12	−.27	.05	−.17
Pupil	−.01	.10	.20	−.05	−.10	−.12	−.10	.04

Note. Given the high accuracies on the SDT, CDT, and LDT, correlations between reaction times, but not accuracies, are shown here. PTA = pure-tone threshold; Rspan = reading span; LM = letter memory; LDT = lexical decision task; TRT = Text Reception Threshold test; Ext. = extraversion; HSPS = Highly Sensitive Person Scale; NASA-TLX = National Aeronautics and Space Administration Task Load Index; CDT = complex dual-task; SDT = semantic dual-task; CSCT = Cognitive Spare Capacity Test; LS = listening span; LP = low predictability; HP = high predictability; RM = running memory.

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

between noise and the individual difference variables. The interaction terms were included to assess whether the influence of the individual difference variables differed as a function of noise condition. For example, the interaction could address the question of whether WM has a stronger relationship with effort in the hard condition than in the easy condition.

Given strong intercorrelations within the subcomponents of the NASA-TLX task, the CSCT, and the LS task, and the similarities in how the subcomponents correlated with cognitive predictors (see Table 6), we built individual models for the composites, rather than subcomponents, of each LE task. This also reduced the number of comparisons, minimizing the false positive rate. For each LE measure, a full model was first created using all the fixed effects (noise, individual difference variables, and interactions). Variables were selectively removed from the model on the basis of their significance levels and contributions to the total sum of squares. After creating reduced models in which each of the cognitive predictor variables contributed significant or marginally significant (ps < .11) unique variance, we ran likelihood ratio tests comparing the full and reduced models to ensure that the reduced model was preferred over the full model (ps ≥ .12; note that the null hypothesis for this comparison is that the coefficients of the variables that differentiate the full from the reduced model are 0. Therefore, a large p value suggests that there is not sufficient evidence to reject the null hypothesis, so the smaller model is preferred. Also note that a higher alpha level was chosen because an alpha level of .05 has been reported to be too conservative for model comparisons; Bursac, Gauss, Williams, & Hosmer, 2008). After building the full model, plots of Cook's distance[1] (Cook, 1977) were visually inspected, and highly influential participants (three or fewer per task) were removed.

The predictors and interactions that were selected for each of the final models are shown in Table 7. When interactions between predictors and noise were significant but the main effects of the predictors were not, the main effects were included in the model but were not reported here. Values shown represent the influence of each predictor if it were added as the final step in the model (thus, the unique variance that each contributes). For all of the LE measures except the CSCT, noise was included as a predictor in the final model with a slope in the expected direction, indicating that louder background noise was associated with more subjectively rated difficulty, slower reaction times, poorer memory performance, and larger pupil dilation.

Scores on all LE measures except pupillometry were correlated with either Rspan or LM (see Table 5, lower), such that greater Rspan and LM scores were associated with faster reaction times, better recall, or lower subjective difficulty ratings on the LE tasks. Given the collinearity between Rspan and LM, only one remained in the mixed-effects models, but all models predicting subjective and behavioral measures of LE included one of the two, indicating that these tasks all rely on WM to some extent. Recall that performance on the CSCT is typically not correlated with Rspan in noise but is correlated with LM performance. Although both correlations were found initially, the mixed-effects model suggests that LM contributes additional unique variance beyond the contributions of Rspan. Thus, given that the Rspan task requires deeper processing than the LM task, it appears that performance on the CSCT relies more heavily on the storage component of WM than on the processing component.

Both dual-task measures of LE (SDT and CDT) were significantly positively related to performance on the

[1] Cook's distance is a measure of a data point's "influence" that quantifies the effect of removing each individual data point on the least-squares regression line.

Table 7. Estimates and *t* values for all cognitive and personality predictors that were significant or marginally significant ($p < .11$) in the final model for each listening effort measure.

Task	Noise	Rspan	LM	LDT	Simon	TRT	Extraversion	HSPS
NASA-TLX	β = .71 (.04) $t = 17.65^{***}$	β = –.23 (.08) $t = –2.78^{**}$						β = .34 (.08) $t = 4.25^{***}$
NASA-TLX interaction with noise		**β = .11 (.04)** $\mathbf{t = 2.92}^{**}$			**β = .07 (.04)** $\mathbf{t = 1.69}^{\dagger}$			
CDT	β = .19 (.03) $t = 5.76^{***}$		β = –.14 (.08) $t = –1.67^{\dagger}$	β = .36 (.09) $t = 4.05^{***}$				β = .30 (.08) $t = 3.55^{***}$
CDT interaction with noise			**β = –.07 (.03)** $\mathbf{t = –2.17}^{*}$		**β = –.06 (.03)** $\mathbf{t = –1.78}^{\dagger}$			**β = .06 (.03)** $\mathbf{t = 1.93}^{\dagger}$
SDT	β = .30 (.05) $t = 5.90^{***}$		β = –.20 (.09) $t = –2.22^{*}$	β = .42 (.10) $t = 4.29^{***}$		β = .17 (.08) $t = 2.02^{*}$		
SDT interaction with noise								
CSCT			β = .30 (.11) $t = 2.80^{**}$					
CSCT interaction with noise		**β = .21 (.11)** $\mathbf{t = 1.94}^{\dagger}$					**β = –.19 (.08)** $\mathbf{t = –2.26}^{*}$	
LS	β = –.31 (.06) $t = –5.23^{***}$	β = .46 (.10) $t = 4.84^{***}$		β = –.15 (.07) $t = –2.12^{*}$		β = –.20 (.08) $t = –2.56^{*}$		
LS interaction with noise		**β = –.17 (.07)** $\mathbf{t = –2.43}^{*}$	**β = .20 (.07)** $\mathbf{t = 2.87}^{**}$					
RM	β = –.58 (.13) $t = –4.42^{***}$		β = .19 (.08) $t = 2.35^{*}$			β = –.16 (.08) $t = –2.16^{*}$		
RM interaction with noise				**β = .29 (.15)** $\mathbf{t = 1.90}^{\dagger}$				
Pupil	β = .26 (.07) $t = 3.81^{***}$							
Pupil interaction with noise					**β = .16 (.07)** $\mathbf{t = 2.43}^{*}$			

Note. Pure-tone threshold was not a significant predictor in any of the final models for any measure of listening effort, so it is not included here. Significant interactions between predictor variables and noise are shown in the bold rows. NASA-TLX = National Aeronautics and Space Administration Task Load Index; CDT = complex dual-task; SDT = semantic dual-task; CSCT = Cognitive Spare Capacity Test; LS = listening span; RM = running memory; Rspan = reading span; LM = letter memory; LDT = lexical decision task; TRT = Text Reception Threshold test; HSPS = Highly Sensitive Person Scale.

$^{\dagger}p < .11$. $^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

LDT in the mixed-effects models, suggesting that participants with faster lexical decision latencies were also faster at the speeded dual-task judgments. LDT latencies were negatively associated with performance on the LS task, indicating that participants with faster processing speeds performed better on the word recall task. This finding is consistent with previous research assessing the relationships between processing speed, WM, and fluid intelligence (Fry & Hale, 1996, 2000).

TRT score emerged as a significant predictor for the SDT, LS, and RM tasks, such that people with lower TRTs (those who were able to decipher the written text even when it was highly masked) had faster reaction times and better recall than those with higher thresholds. The link between LS and TRT has been demonstrated previously (Besser et al., 2013). Some prior studies have found that better TRT scores are associated with greater peak pupil dilation in listeners with normal hearing (Koelewijn, Zekveld, Festen, Rönnberg, et al., 2012; Zekveld & Kramer, 2014; Zekveld et al., 2011), but this may only occur when the masking noise is interfering speech (unlike the steady-state noise used here) or when the SNR is individualized. Thus, the lack of relationship between TRT and peak pupil dilation is not surprising. This is the first study to demonstrate a link between measures of LE and SPS. People with higher HSPS scores rated the tasks as subjectively more difficult and had slower CDT latencies.

The models also included a total of 11 interactions between cognitive or personality predictors and noise condition. About half of these (n = 6) showed that better performance on the cognitive measures (faster reaction times, higher spans, lower TRT scores) was more strongly correlated with scores on the LE measure in the hard condition than in the easy condition. For example, the positive interaction between LM and LS revealed that there is a stronger relationship between LM and LS in the hard condition than in the easy condition; people with higher LM scores tended to show smaller changes in LS performance when moving from the easy task to the hard task. The patterns were similar for the interactions between noise and Simon score (for the NASA-TLX and pupil task), LM score (for the CDT), and Rspan score (for the CSCT). In addition, the marginally significant (p = .057) HSPS × Noise interaction indicates that the relationship between HSPS score and CDT latency is stronger in the hard than in the easy condition.

In contrast, the remainder of the significant interactions (n = 5) showed that better performance on memory and reaction time measures was more strongly related to scores on LE measures in the easy than in the hard condition. For example, higher Rspan scores were associated with lower NASA-TLX difficulty ratings overall (main effect), but that was particularly true in the easy condition. Thus, the NASA-TLX difficulty ratings of participants with higher Rspan scores tended to be more affected by noise. This was also the case for the interactions between noise and CDT (for the Simon task), Rspan score (for the LS task), and LDT score (for the RM task). In addition, extraversion showed a stronger positive relationship with CSCT performance in the easy than in the hard condition.

Discussion

Research Aims

Aim 1: Convergent Validity

The correlations among measures of LE varied considerably, with the strength ranging from r = .01 to r = .53 for different task pairs. The average of the absolute values of the correlations among different measures of LE (excluding the correlations among subcomponents of the same task) was r = .22, indicating a relatively weak overall relationship between performance on the multiple measures of LE. Weak correlations among subjective and behavioral measures have been well-documented (cf., Downs & Crum, 1978; Johnson et al., 2015; Seeman & Sims, 2015), but these results demonstrate relatively modest correlations among the behavioral measures as well, with the average absolute value of correlations in Table 5 (excluding the subjective and physiological measures and subcomponents of the same task) being r = .30. However, nearly all of the correlations among the measures were in the expected direction, indicating that better performance on one LE measure (faster reaction time, better recall) tended to be associated with better performance on others, although many comparisons did not reach significance. A notable exception is that scores from the pupillometry measure were not significantly correlated with performance on any of the other tasks. Although pupil dilation was sensitive to changes in noise, it appears to be measuring a different underlying construct from the other measures of LE (see the discussion of Aim 3 for more on this issue).

The analyses also revealed that the consequences of additional noise differ for some pairs of tasks. Although the effects of noise were most consistent within the recall measures, some significant interactions still emerged within these tasks, indicating that changing from the easy to the hard condition affected task pairs differently (e.g., LS, HP and RM). Therefore, it is not always the case that an individual who is impaired by noise on one task is impaired by noise on another task. Future research should evaluate the roots of individual differences in how noise affects task performance. For example, perhaps some participants are better able to make use of semantic context, and therefore show small impairments from the addition of noise in the LS, HP task. In conditions without semantic context, such as LS, LP and RM, drops in performance with the addition of noise are more consistent.

Aim 2: Sensitivity

Although only a few papers have evaluated the sensitivities of LE measures to changes in task difficulty, the results of this study support previous reports. Our results support and extend the Picou and Ricketts (2014) finding that the SDT is more sensitive than the CDT in detecting changes in LE between quiet and noisy conditions. Here, we show that the SDT is also more sensitive than the CDT under conditions in which the SNR changes from easy to hard. Thus, the SDT is sensitive to subtle changes in noise level, suggesting that it can be effectively used as a measure of LE under a variety of noise conditions.

We found that the LE measure that was numerically most sensitive to changes in SNR was the NASA-TLX subjective measure. These results are consistent with those demonstrated by Seeman and Sims (2015), who found that the NASA-TLX self-report measure was more sensitive to changes in SNR than physiological and behavioral measures. Similarly, Johnson et al. (2015) found that the NASA-TLX measure was more sensitive to changes in SNR than a recall measure comparable to the LS task employed in this experiment. However, given that subjective ratings of LE are often correlated with speech recognition performance (Downs & Crum, 1978) and not with objective measures of LE (Johnson et al., 2015; Seeman & Sims, 2015), these results should not be taken as evidence that subjective measures are the most effective measures of LE; under conditions that produce similar performance, subjective measures may fail to detect the changes in LE that have been identified using objective measures (Downs & Crum, 1978; Sarampalis et al., 2009). Thus, although they were the most sensitive to changes in SNR, subjective measures of LE may not be as versatile as other measures of LE.

Overall, it appears that other than subjective LE measures, recall measures and the SDT are most sensitive to changes in SNR, suggesting that these tasks require deep processing and many cognitive resources. That is, under difficult listening conditions, more resources must be devoted to recognizing speech, leaving fewer resources available to complete difficult secondary or recall tasks (see Picou & Ricketts, 2014, for more on this). In contrast, tasks that require more automatic processing, like the CDT, can be completed with fewer cognitive resources, so performance is less affected by changes in SNR. However, deeper processing cannot solely account for differences in task sensitivity, as the RM task, which is primarily an updating task and requires little processing, showed numerically higher sensitivity than both the LS task and the SDT. One possible explanation for this is that the words used in the RM task tended to be longer and less common than those of other tasks, which may have increased processing load. Future studies should assess how lexical characteristics of stimuli contribute to LE.

The larger effect size for the LP portion of the LS task compared with the HP portion may suggest that processing semantically meaningful sentences requires fewer cognitive resources than processing unpredictable sentences, so differences between easy and hard listening conditions are less apparent. It should be noted that these results are inconsistent with those in Johnson et al. (2015), in which larger effect sizes were observed in HP contexts compared with LP contexts. Therefore, future research should evaluate the conditions under which semantic context reduces LE.

Aim 3: Individual Differences

PTA was not significantly correlated with any of the LE measures and was not retained by any of the models. Although people with hearing loss tend to expend more LE than individuals with normal hearing (Bourland-Hicks & Tharpe, 2002), our results suggest that the moderate variability in PTA among young adults with normal hearing is not sufficient to systematically affect performance on LE tasks. All main effects associated with cognitive predictors indicated that greater cognitive capacity (higher Rspan and LM scores, faster processing speeds, smaller incongruency costs, and lower TRT scores) was associated with better scores on the LE measures. These results suggest that greater cognitive capacity is associated with decreased LE. The most robust predictors of performance on LE tasks were Rspan and LM scores (which were significantly correlated with each other as well). Performances on the CSCT, the LS task, and to a lesser degree, the RM task were all significantly correlated with these memory measures, indicating that successful completion of the LE tasks relies on similar cognitive mechanisms to those recruited for these commonly used WM and updating tasks.

The interactions between noise and the cognitive and personality measures were included in the models to determine whether these variables are stronger predictors of LE in challenging listening situations, in line with the predictions of the ELU. The majority of the possible interactions we entered into the model were not significant (of 35 possible interactions among nonpersonality variables and 49 interactions total, only 11 emerged as marginal or significant), and those that were rendered conflicting results—some showed stronger relationships between cognitive predictors and LE measures in the hard condition, and some showed stronger relationships in the easy condition. Therefore, the current study does not provide support for the claim that cognitive capacity systematically affects processing load more in challenging listening situations than in easy ones. Thus, the results are more in line with the prediction of the cognitive efficiency hypothesis that greater cognitive capacity is associated with decreased processing load, regardless of the difficulty of the listening situation.

Because of the large number of tasks we included in this study, we only tested the LE measures at two SNRs. Given this limited range of difficulty, it is possible that the hard noise condition we included (SNR = −2 dB) was not sufficiently difficult to show that cognitive capacity decreases effort more in taxing situations (see Wu et al., 2016 and Zekveld & Kramer, 2014 for more on how SNR affects performance on LE tasks). Indeed, word recognition accuracy in the hard condition was still relatively high (70%–81%, see Table 3). Future studies that include fewer separate tasks should consider including a larger range of SNRs to assess whether greater task difficulty allows effects consistent with the predictions of the ELU hypothesis, namely that cognitive capacity has a larger effect in more difficult listening conditions, to emerge.

Additional Considerations and Recommendations

Effects of Noise

Future studies should assess whether and how the results of this study differ for a variety of difficulty manipulations and for other types of noise. We used changes in SNR to manipulate difficulty because it could easily be applied to all of the LE tasks, but a study involving fewer LE tasks could assess difficulty in other ways (e.g., amount of lexical competition, semantic predictability, etc.). The results may also differ for different forms of background noise; for example, Mishra et al. (2013a) found that differences in CSC between audio-only and audiovisual speech only emerged in some types of noise. Thus, future work should assess how different types of noise or noise demands (e.g., informational vs. energetic masking) affect the effort necessary to process speech.

This study is the first to examine how CSC is affected by changes in SNR (i.e., changing from an easy SNR to a hard SNR rather than from quiet to noise). The CSCT did not show any significant effects of noise, indicating that CSC does not differ systematically between the easy and hard listening conditions used here. Future studies should address whether the CSCT is sensitive to larger differences in SNR between conditions. Both dual-task measures were affected by changes in task difficulty, but the SDT was numerically more sensitive than the CDT. Thus, researchers designing LE experiments using dual-task paradigms should carefully consider expected effect sizes before deciding which task to use.

The bulk of work on LE (e.g., Desjardins & Doherty, 2014; Fraser et al., 2010; Gosselin & Gagné, 2011a, 2011b; Picou & Ricketts, 2014) has presented speech at a consistent level and manipulated SNR by changing the amplitude of the masking noise, so the current study followed this convention. However, an anonymous reviewer pointed out that it is possible that decreases in performance on LE tasks associated with increased noise might be caused by noise affecting performance on the cognitive task, rather than (or in addition to) the speech task. For example, it may be that the ability to make “noun” judgments in the SDT would be adversely affected by noise, even if words were visually presented. Future work should assess whether and how background noise might affect the cognitive tasks, above and beyond increasing LE.

Implementation of LE Tasks

A key feature on which LE tasks differ is whether they produce separate metrics for intelligibility and effort. An attractive feature of the dual-task paradigms, for example, is that performance on the word identification task is measured independently from reaction times to the secondary task. In contrast, in the CSCT and RM tasks, it is difficult to distinguish between performance on the task and effort because incorrectly perceiving an item may be interpreted as failing to correctly recall it. Researchers should consider whether the purposes of their studies would benefit from clearly dissociable measures of effort and performance. An additional consideration for researchers implementing the CSCT is that it was also the task on which the most participants reported giving up; even with a shortened presentation, participants anecdotally reported that the task was very difficult to complete.

Distinguishing Between LE and Related Constructs

Pupil dilation was not correlated with any measure of cognitive ability or personality trait in this study. There was a significant Simon × Noise interaction in the pupillometry model, such that participants with low Simon scores (those who showed smaller incongruency costs and thus better inhibition) tended to be less affected by the addition of noise than those with higher Simon scores. However, this effect was relatively small, so of the available LE measures, the pupillometry task appears to rely the least on cognitive and personality variables. However, it is also important to note that the listening portion of the pupillometry task may have been less cognitively demanding than other tasks, which may influence the extent to which an effect of LE is detected. Given other research showing a link between pupillometry and cognitive load (Unsworth & Robison, 2015), future work should explore how changing task demands in pupillometry tasks affects the amount of LE required.

This study is the first to demonstrate that individuals who score higher on sensory sensitivity show increased effort as measured by some LE tasks. This adds to work that has found that high SPS is associated with greater perceived stress and more frequent symptoms of ill health (Benham, 2006) and suggests that individuals with higher SPS may also be more susceptible to the detrimental effects of difficult listening situations. These results cannot be solely attributed to participants with high SPS tending to self-report higher values on both stress and effort, as HSPS score also emerged as a significant predictor in the model for a behavioral task, the CDT.

Our results also suggest that researchers should think carefully about whether to include tests of cognitive function along with their measures of LE in order to statistically control for individual difference variables. For example, a study assessing how aging affects LE may want to consider including the Rspan or LM task to help dissociate how aging affects general cognitive abilities like WM and updating from how aging affects the effort necessary to understand speech (see Smith, Pichora-Fuller, & Alexander, 2016).

The current study examined the relationship between task demands and effort. However, another important factor that may influence performance is participants’ motivation to complete the task successfully (Kahneman, 1973; Peelle, 2017; Pichora-Fuller et al., 2016). Although our exclusion of participants who reported high levels of giving up is a way of removing participants who appear to be unmotivated to complete a particular task, future work should try to quantify motivation more systematically. Indeed, motivation may modulate the extent to which cognitive variables predict LE; simply having greater cognitive capacity does not mean that an individual will allocate it during speech processing (see Smith & Pichora-Fuller, 2015). Pichora-Fuller et al. (2016) provided a clear framework for predicting the relationship between effort, task demand, and motivation that future research may use as a guide (see also Peelle, 2017).

Conclusions

The LE literature has relied on a wide range of subjective, behavioral, and physiological measures to assess the effort necessary to successfully recognize speech. Our results indicate that it is inadvisable to draw conclusions across studies that use different LE tasks, given that the measures do not show consistent strong intercorrelations, and they differ in their relationships with well-established cognitive and personality predictors. However, all relationships between cognitive variables and LE measures indicated that better cognitive ability was associated with reduced effort. These findings add to the existing literature showing that cognitive ability improves performance on speech processing tasks and suggest that, in addition to enhancing recognition, cognitive ability reduces the deleterious effects of additional noise.

Acknowledgments

Carleton College supported this work. The authors are grateful to Janna Wennberg, Maryam Hedayati, Annie Zanger, Lucia Ray, and Sasha Mayn for helpful conversations and assistance with data collection; to Hettie Stern and Lydia Ding for their input during task selection; to Mitch Sommers for recommendations about experiment design; to Nathan Rose for guidance on cognitive measures and feedback on an earlier draft; to Laura Chihara and Andrew Poppick for statistical guidance; to Erin Picou for sharing stimulus lists for the CDT and SDT paradigms; to Gitte Keidser and Mary Rudner for sharing stimulus lists for the CSCT; to the Engle lab for providing RSpan sentences; and to Julie Neiworth for feedback on the article.

References

Adobe Audition [Computer software]. (2017). Retrieved from https://www.adobe.com/products/audition.html

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205*(4412), 1289–1292.

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47*(Suppl. 2), S53–S71.

Alhanbali, S., Dawes, P., Lloyd, S., & Munro, K. J. (2017). Self-reported listening-related effort and fatigue in hearing-impaired adults. *Ear and Hearing, 38*(1), e39–e48.

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision, 14*(4). Online only publication. https://doi.org/10.1167/14.4.1

Aron, E. N., & Aron, A. (1997). Sensory-processing sensitivity and its relation to introversion and emotionality. *Journal of Personality and Social Psychology, 73*(2), 345–368.

Aron, E. N., Aron, A., & Jagiellowicz, J. (2012). Sensory processing sensitivity: A review in the light of the evolution of biological responsivity. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc, 16*(3), 262–282.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91*(2), 276–292.

Benham, G. (2006). The highly sensitive person: Stress and physical symptom reports. *Personality and Individual Differences, 40*(7), 1433–1440.

Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2013). How linguistic closure and verbal working memory relate to speech recognition in noise—A review. *Trends in Amplification, 17*(2), 75–93.

Besser, J., Zekveld, A. A., Kramer, S. E., Rönnberg, J., & Festen, J. M. (2012). New measures of masked text recognition in relation to speech-in-noise perception and their associations with age and cognitive abilities. *Journal of Speech, Language, and Hearing Research, 55*(1), 194–209.

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research, 27*(1), 32–48.

Boothroyd, A., & Schauer, A. (2015). Lowest acceptable performance level. In *International Collegium of Rehabilitative Audiology*. Berkeley, CA. https://doi.org/10.13140/RG.2.1.1061.0324

Borg, G. (1990). Psychophysical scaling with applications in physical work and the perception of exertion. *Scandinavian Journal of Work, Environment and Health, 16*(Suppl. 1), 55–58.

Bourland-Hicks, C., & Tharpe, A. M. (2002). Listening effort and fatigue in school-age children with and without hearing loss. *Journal of Speech, Language, and Hearing Research, 45*(3), 573–584.

Brault, L. M., Gilbert, J. L., Lansing, C. R., McCarley, J. S., & Kramer, A. F. (2010). Bimodal stimulus presentation and expanded auditory bandwidth improve older adults' speech perception. *Human Factors, 52*(4), 479–491.

Broadbent, D. E. (1958). The effects of noise on behavior. In D. E. Broadbent (Ed.), *Perception and communication* (pp. 81–107). Elmsford, NY: Pergamon.

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.

Bunting, M., Cowan, N., & Saults, J. S. (2006). How does running memory span work? *The Quarterly Journal of Experimental Psychology, 59*(10), 1691–1700.

Bursac, Z., Gauss, C. H., Williams, D. K., & Hosmer, D. W. (2008). Purposeful selection of variables in logistic regression. *Source Code for Biology and Medicine, 3*(1), 17.

Clark-Carter, D. (2009). *Quantitative psychological research: The complete student's companion* (3rd ed.). Hove, United Kingdom: Psychology Press.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review, 12*(5), 769–786.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences, 19*(1), 15–18.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450–466.

Desjardins, J. L., & Doherty, K. A. (2013). Age-related changes in listening effort for various types of masker noises. *Ear and Hearing, 34*(3), 261–272.

Desjardins, J. L., & Doherty, K. A. (2014). The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. *Ear and Hearing, 35*(6), 600–610.

Downs, D. W. (1982). Effects of hearing aid use on speech discrimination and listening effort. *Journal of Speech and Hearing Disorders, 47*(2), 189–193.

Downs, D. W., & Crum, M. A. (1978). Processing demands during auditory learning under degraded listening conditions. *Journal of Speech and Hearing Research, 21*(4), 702–714.

Eysenck, H. J. (1967). *The biological basis of personality*. Piscataway, NJ: Transaction.

Eysenck, H. J. (1990). Biological dimensions of personality. In L. A. Pervin (Ed.), *Handbook of personality: Theory and research* (pp. 244–276). New York, NY: Guilford.

Feuerstein, J. F. (1992). Monaural versus binaural hearing: Ease of listening, word recognition, and attentional effort. *Ear and Hearing, 13*(2), 80–86.

Foo, C., Rudner, M., Rönnberg, J., & Lunner, T. (2007). Recognition of speech in noise with new hearing instrument compression release settings requires explicit cognitive storage and processing capacity. *Journal of the American Academy of Audiology, 18*(7), 618–631.

Frank, A. (2014). *mer-utils.R*. Retrieved from https://github.com/aufrank/R-hacks/blob/master/mer-utils.R

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, 53*(1), 18–33.

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science, 7*(4), 237–241.

Fry, A. F., & Hale, S. (2000). Relationships among processing speed, working memory, and fluid intelligence in children. *Biological Psychology, 54*(1–3), 1–34.

Füllgrabe, C., & Rosen, S. (2016). On the (un)importance of working memory in speech-in-noise processing for listeners with normal hearing thresholds. *Frontiers in Psychology, 7,* 1268.

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.

Geen, R. G. (1984). Preferred stimulation levels in introverts and extroverts: Effects on arousal and performance. *Journal of Personality and Social Psychology, 46*(6), 1303–1312.

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

Gosselin, P. A., & Gagné, J.-P. (2011b). Older adults expend more listening effort than young adults recognizing speech in noise. *Journal of Speech, Language, and Hearing Research, 54*(3), 944–958. https://doi.org/10.1044/1092-4388(2010/10-0069)

Guerreiro, M. J. S., Murphy, D. R., & Van Gerven, P. W. M. (2013). Making sense of age-related distractibility: The critical role of sensory modality. *Acta Psychologica, 14*(2), 184–194.

Hällgren, M., Larsby, B., Lyxell, B., & Arlinger, S. (2005). Speech understanding in quiet and noise, with and without hearing aids. *International Journal of Audiology, 44*(10), 574–583.

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.

Howard, C. S., Munro, K. J., & Plack, C. J. (2010). Listening effort at signal-to-noise ratios that are typical of the school classroom. *International Journal of Audiology, 49*(12), 928–932.

Hughes, K. C., & Galvin, K. L. (2013). Measuring listening effort expended by adolescents and young adults with unilateral or bilateral cochlear implants or normal hearing. *Cochlear Implants International, 14*(3), 121–129.

Humes, L. E. (2007). The contributions of audibility and cognitive factors to the benefit provided by amplified speech to older adults. *Journal of the American Academy of Audiology, 18*(7), 590–603.

Jagiellowicz, J., Xu, X., Aron, A., Aron, E., Cao, G., Feng, T., & Weng, X. (2011). The trait of sensory processing sensitivity and neural responses to changes in visual scenes. *Social Cognitive and Affective Neuroscience, 6*(1), 38–47.

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.

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

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583–1585.

Kalikow, D. N., Stevens, K. N., & Elliott, L. L. (1977). Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *The Journal of the Acoustical Society of America, 61*(5), 1337–1351.

Keidser, G., Best, V., Freeston, K., & Boyce, A. (2015). Cognitive spare capacity: Evaluation data and its association with comprehension of dynamic conversations. *Frontiers in Psychology, 6,* 597.

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.

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.

Koelewijn, T., Zekveld, A. A., Festen, J. M., & Kramer, S. E. (2012). Pupil dilation uncovers extra listening effort in the presence of a single-talker masker. *Ear and Hearing, 33*(2), 291–300.

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., & Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *International Journal of Otolaryngology, 2012,* 865731.

Kramer, S. E., Kapteyn, T. S., Festen, J. M., & Kuik, D. J. (1997). Assessing aspects of auditory handicap by means of pupil dilation. *Audiology: Official Organ of the International Society of Audiology, 36*(3), 155–164.

Kramer, S. E., Kapteyn, T. S., & Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for

Hearing and Work. *International Journal of Audiology, 45*(9), 503–512.

Kryter, K. D. (1970). *The effects of noise on man*. Cambridge, MA: Academic Press.

Kuchinsky, S. E., Ahlstrom, J. B., Vaden, K. I., Jr., Cute, S. L., Humes, L. E., Dubno, J. R., & Eckert, M. A. (2013). Pupil size varies with word listening and response selection difficulty in older adults with hearing loss. *Psychophysiology, 50*(1), 23–34.

Larsby, B., Hällgren, M., Lyxell, B., & Arlinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: Effects of different noise backgrounds in normal-hearing and hearing-impaired subjects. *International Journal of Audiology, 44*(3), 131–143.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *International Journal of Audiology, 42*(Suppl. 1), S49–S58. https://doi.org/10.3109/14992020309074624

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.

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7–8), 953–978.

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: Human Experimental Psychology, 58A*(1), 22–33.

McGarrigle, R., Dawes, P., Stewart, A. J., Kuchinsky, S. E., & Munro, K. J. (2016). Pupillometry reveals changes in physiological arousal during a sustained listening task. *Psychophysiology, 54*(2), 193–203.

McGarrigle, R., Munro, K. J., Dawes, P., Stewart, A. J., Moore, D. R., Barry, J. G., & Amitay, S. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *International Journal of Audiology, 53*(7), 433–445.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology, 90*(2), 227–234.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013a). Seeing the talker's face supports executive processing of speech in steady state noise. *Frontiers in Systems Neuroscience, 7,* 96.

Mishra, S., Lunner, T., Stenfelt, S., Rönnberg, J., & Rudner, M. (2013b). Visual information can hinder working memory processing of speech. *Journal of Speech, Language, and Hearing Research, 56,* 1120–1132.

Mishra, S., Stenfelt, S., Lunner, T., Rönnberg, J., & Rudner, M. (2014). Cognitive spare capacity in older adults with hearing loss. *Frontiers in Aging Neuroscience, 6,* 96.

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49–100.

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.

Murphy, D. R., Craik, F. I. M., Li, K. Z. H., & Schneider, B. A. (2000). Comparing the effects of aging and background noise on short-term memory performance. *Psychology and Aging, 15*(2), 323–334.

Ng, E. H. N., Rudner, M., Lunner, T., Pedersen, M. S., & Rönnberg, J. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *International Journal of Audiology, 52*(7), 433–441.

Pals, C., Sarampalis, A., & Baskent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech, Language, and Hearing Research, 56*(4), 1075–1084.

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116*(2), 220–244.

Peelle, J. E. (2017). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear and Hearing, 39*(2), 204–214. https://doi.org/10.1097/AUD.0000000000000494

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., . . . 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.

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.

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. Y. (2013). How hearing aids, background noise, and visual cues influence objective listening effort. *Ear and Hearing, 34*(5), e52–e64.

Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology, 47*(3), 560–569.

Pittman, A. (2011). Children's performance in complex listening conditions: Effects of hearing loss and digital noise reduction. *Journal of Speech, Language, and Hearing Research, 54*(4), 1224–1239.

Pratte, M. S., Rouder, J. N., Morey, R. D., & Feng, C. (2010). Exploring the differences in distributional properties between Stroop and Simon effects using delta plots. *Attention, Perception and Psychophysics, 72*(7), 2013–2025.

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

Rakerd, B., Seitz, P. F., & Whearty, M. (1996). Assessing the cognitive demands of speech listening for people with hearing losses. *Ear and Hearing, 17*(2), 97–106.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment: Official Organ of the European Association of Psychological Assessment, 28*(3), 164–171.

Rönnberg, J. (2003). Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: A framework and a model. *International Journal of Audiology, 42,* S68–S76.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., . . . Rudner, M. (2013). The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience, 7,* 31.

Rönnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology, 47*(Suppl. 2), S99–S105.

Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., . . . Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, 17*(3), 225–246.

Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnberg, J. (2012). Working memory capacity may influence perceived effort during aided speech recognition in noise. *Journal of the American Academy of Audiology, 23*(8), 577–589.

Rudner, M., Ng, E. H. N., Ronnberg, N., Mishra, S., Ronnberg, J., Lunner, T., & Stenfelt, S. (2011). Cognitive spare capacity as a measure of listening effort. *Journal of Hearing Science, 1*(2), EA47–EA49. Retrieved from http://www.divaportal.org/smash/record.jsf?pid=diva2%3A792812&dswid=640

Rudner, M., Rönnberg, J., & Lunner, T. (2011). Working memory supports listening in noise for persons with hearing impairment. *Journal of the American Academy of Audiology, 22*(3), 156–167.

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), 1230–1240.

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

Simon, J. R. (1969). Reactions toward the source of stimulation. *Journal of Experimental Psychology, 81*(1), 174–176.

Smith, S. L., & Pichora-Fuller, M. K. (2015). Associations between speech understanding and auditory and visual tests of verbal working memory: Effects of linguistic complexity, task, age, and hearing loss. *Frontiers in Psychology, 6,* 1394.

Smith, S. L., Pichora-Fuller, M. K., & Alexander, G. (2016). Development of the word auditory recognition and recall measure: A working memory test for use in rehabilitative audiology. *Ear and Hearing, 37*(6), e360–e376.

Smolewska, K. A., McCabe, S. B., & Woody, E. Z. (2006). A psychometric evaluation of the Highly Sensitive Person Scale: The components of sensory-processing sensitivity and their relation to the BIS/BAS and "Big Five." *Personality and Individual Differences, 40*(6), 1269–1279.

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–68S.

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.

Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

Tobii Studio Professional (Version 3.2) [Computer software]. (2013). Retrieved from https://www.tobiipro.com/product-listing/tobii-pro-studio/

Tun, P. A., McCoy, S., & Wingfield, A. (2009). Aging, hearing acuity, and the attentional costs of effortful listening. *Psychology and Aging, 24*(3), 761–766.

Unsworth, N., & Robison, M. K. (2015). Individual differences in the allocation of attention to items in working memory: Evidence from pupillometry. *Psychonomic Bulletin and Review, 22*(3), 757–765.

Van Der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., . . . Wartenburger, I. (2010). Resource allocation and fluid intelligence: Insights from pupillometry. *Psychophysiology, 47*(1), 158–169. https://doi.org/10.1111/j.1469-8986.2009.00884.x

Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PloS One, 7*(9), e43753.

Wagner, A. E., Toffanin, P., & Başkent, D. (2016). The timing and effort of lexical access in natural and degraded speech. *Frontiers in Psychology, 7,* 398.

Wendt, D., Dau, T., & Hjortkjær, J. (2016). Impact of background noise and sentence complexity on processing demands during sentence comprehension. *Frontiers in Psychology, 7,* 345.

Wingfield, A. (2016). Evolution of models of working memory and cognitive resources. *Ear and Hearing, 37*(Suppl. 1), 35S–43S.

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

Wu, Y.-H., Aksan, N., Rizzo, M., Stangl, E., Zhang, X., & Bentler, R. (2014). Measuring listening effort: Driving simulator versus simple dual-task paradigm. *Ear and Hearing, 35*(6), 623–632.

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.

Zekveld, A. A., George, E. L. J., Kramer, S. E., Goverts, S. T., & Houtgast, T. (2007). The development of the Text Reception Threshold test: A visual analogue of the speech reception threshold test. *Journal of Speech, Language, and Hearing Research, 50*(3), 576–584.

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.

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.

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., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing, 32*(4), 498–510.