




Understanding Speech amid the Jingle and Jangle: Recommendations for Improving Measurement Practices in Listening Effort Research

Julia F. Strand ^a, Lucia Ray^a, Naseem H. Dillman-Hasso ^a, Jed Villanueva^a
and Violet A. Brown ^b

^aDepartment of Psychology, Carleton College, Northfield, MN, USA; ^bDepartment of Psychological & Brain Sciences, Washington University in St. Louis, St Louis, MO, USA

ABSTRACT

The latent constructs psychologists study are typically not directly accessible, so researchers must design measurement instruments that are intended to provide insights about those constructs. Construct validation—assessing whether instruments measure what they intend to—is therefore critical for ensuring that the conclusions we draw actually reflect the intended phenomena. Insufficient construct validation can lead to the *jingle fallacy*—falsely assuming two instruments measure the same construct because the instruments share a name—and the *jangle fallacy*—falsely assuming two instruments measure different constructs because the instruments have different names. In this paper, we examine construct validation practices in research on *listening effort* and identify patterns that strongly suggest the presence of jingle and jangle in the literature. We argue that the lack of construct validation for listening effort measures has led to inconsistent findings and hindered our understanding of the construct. We also provide specific recommendations for improving construct validation of listening effort instruments, drawing on the framework laid out in a recent paper on improving measurement practices. Although this paper addresses listening effort, the issues raised and recommendations presented are widely applicable to tasks used in research on auditory perception and cognitive psychology.

ARTICLE HISTORY

Received 2 October 2020
Accepted 10 March 2021

KEYWORDS

Listening effort; speech;
measurement; validity

Many psychological phenomena defy direct observation (e.g., implicit bias, memory capacity, loneliness, learning, etc.), so in order to study them, psychologists must create measures that indirectly measure the latent constructs of interest. The conclusions we can draw from studies using these measures are only sound to the extent that the measures do in fact represent the construct of interest. Evaluating the validity of psychological measurement tools—that is, the extent to which tests measure what they purport to measure (Cohen & Swerdlik, 2010)—has long been recognized as a crucial component of research (e.g., Cronbach & Meehl, 1955). However, it has become increasingly clear in recent years that thoroughly assessing the validity evidence of psychological measures is not common practice. For example, in the field of health education and behavior, 40–93%

CONTACT Julia F. Strand  jstrand@carleton.edu  Department of Psychology, Carleton College, Northfield, MN, USA

© 2021 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

of studies in seven journals included no evidence about validity (Barry, Chaney, Piazza-Gardner, & Chavarria, 2014). In a recent review of scales used in the *Journal of Personality and Social Psychology* in 2014, approximately 46% did not include any evidence of previous validation, and about 19% of those that did not include a citation reported no psychometric evidence whatsoever (Flake, Pek, & Hehman, 2017).

Much of the focus on measurement issues in recent years has centered around survey instruments or clinician evaluations. However, the behavioral tasks used in cognitive psychology are not immune to the consequences of weak validation, though they have received less attention. In cognitive tasks, participants may be asked to respond to a prompt as quickly as possible, recall previously presented information, or make a judgment about a stimulus. In these cases, the response times, recall rates, or other forms of responses are assumed to represent something specific about the mental processes necessary to complete the tasks. However, no matter how precisely response times or error rates are measured, the assumptions about what the tasks represent may be incorrect. In one classic example (Sperling, 1960), participants were shown a 3×3 matrix of letters and then asked to freely recall as many of them as possible. Limits on recall (typically four to five items) were thought to reflect the constraints on what could be encoded during a brief presentation. However, if participants were prompted to recall just one row—and those instructions came following the disappearance of the stimuli—they could consistently recall the entire row, meaning they had access to all nine numbers. This suggests that responses on the free recall task were not measuring the amount of information participants could hold in memory during encoding, but the amount they could report before the memory trace faded. Thus, the measurement instrument was assessing something other than the intended construct.

In this paper, we address the insufficient consideration of validity in cognitive psychology with respect to a specific construct: *listening effort*. We argue that the lack of attention to validity has led to inconsistency in the literature and difficulty forming theoretical frameworks about the cognitive processes underlying listening effort. Finally, we offer recommendations for ways that researchers may increase transparency and evaluate the validity of the measures they use. Although the paper uses listening effort as a running example because it is the authors' area of expertise, the issues raised here are likely applicable to many research areas in cognitive psychology and beyond.

Listening Effort

Listening effort can be defined as “the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task” (Pichora-Fuller et al., 2016). This construct is of interest to researchers seeking to understand the mechanisms underlying human speech perception as well as to clinicians; not only can effortful listening lead to increased anxiety and social isolation (Pichora-Fuller, 2016), but an understanding of listening effort may also be useful for audiologists when assessing intervention plans, such as deciding between competing hearing-aid algorithms (McGarrigle et al., 2014; Sarampalis, Kalluri, Edwards, & Hafter, 2009). Listening effort is distinct from accuracy at identifying speech—indeed, word identification accuracy can be stable despite changes in effort (Sarampalis et al., 2009; Strand, Brown, & Barbour, 2020). Listening effort appears to be higher in noisy listening

conditions than in quiet (Larsby, Hällgren, Lyxell, & Arlinger, 2005; Zekveld, Kramer, & Festen, 2010), for individuals with hearing loss than for individuals with normal hearing (Bourland-Hicks & Tharpe, 2002), and for older relative to young adults (Desjardins & Doherty, 2013; Tun, McCoy, & Wingfield, 2009).

Various models and frameworks have been proposed to describe the mechanism underlying listening effort. The Ease of Language Understanding model (Rönnerberg et al., 2013; Rönnerberg, Rudner, Foo, & Lunner, 2008) proposes that speech processing is automatic when the perceptual input matches a representation in long term memory, but when the speech input is degraded and therefore does not match a representation, explicit processing resources (e.g., working memory) are recruited to facilitate listening. This explicit processing that occurs in difficult listening conditions is assumed to be slower and more effortful than the implicit processing that occurs in pristine conditions. The Framework for Understanding Effortful Listening (Pichora-Fuller et al., 2016) builds on this theory by considering how other factors, including motivation, influence listening effort.

Although listening effort is an intuitive concept, there is no consensus about the most appropriate way to measure it. More than 24 different tasks have been used to measure listening effort (Strand, Brown, Merchant, Brown, & Smith, 2018), generally divided into three major classes (see Gagné, Besser, & Lemke, 2017a; McGarrigle et al., 2014; Pichora-Fuller et al., 2016): self-report, physiological, and behavioral. Self-report measures ask participants to report the amount of effort they exerted during a listening task (Johnson, Xu, Cox, & Pendergraft, 2015). Physiological measures quantify bodily changes that are assumed to result from the increased arousal associated with effortful listening, such as changes in pupil size (i.e., pupillometry; Zekveld, Kramer, & Festen, 2011) and heart rate variability (Mackersie & Cones, 2011). Behavioral measures rely on overt responses from the participant and are often accomplished using a dual-task paradigm in which participants complete a primary listening task while simultaneously performing a secondary task (Broadbent, 1958; see Gagné et al., 2017a for a review). For example, participants may be asked to listen to speech (the primary task) while simultaneously responding to visually-presented stimuli (Picou & Ricketts, 2014; Sarampalis et al., 2009), identifying tactile patterns (Fraser, Gagné, Alepins, & Dubois, 2010; Gosselin & Gagné, 2011), judging whether words rhyme (Pals, Sarampalis, & Baskent, 2013), or completing other secondary tasks. Slower reaction times or poorer accuracy on the secondary tasks are assumed to represent greater effort (see Gagné, Besser, & Lemke, 2017b for a review of the use of dual-task paradigms to measure effort). Behavioral tasks also include recall paradigms in which participants must hold spoken items in memory for recall later. Recall paradigms may include paired-associate tasks (Picou, Ricketts, & Hornsby, 2011), running memory tasks (McCoy et al., 2005; Sommers & Phelps, 2016), listening span tasks (Johnson et al., 2015; Ng, Rudner, Lunner, Pedersen, & Rönnerberg, 2013; Pichora-Fuller, Schneider, & Daneman, 1995; Sarampalis et al., 2009), or cognitive spare capacity tests (Mishra, Lunner, Stenfelt, Rönnerberg, & Rudner, 2013a, 2013b). The rationale for these tasks is that greater listening effort leads to fewer resources remaining for rehearsal or encoding (Rabbitt, 1968).

The presence of multiple instruments to measure a given construct is not necessarily cause for concern. However, in the listening effort literature, evidence is mounting that these various measures are not interchangeable. For example, even within the same

participants and using similar stimuli, different listening effort tasks are often not (or only weakly) correlated, indicating they may not be tapping into the same underlying construct (Alhanbali, Dawes, Millman, & Munro, 2019; Strand et al., 2018). Measures also differ in how sensitive they are to changes in the level of background noise, and impairment by the addition of noise in one task does not necessarily predict impairment on another task (Strand et al., 2018). Finally, two experiments that use the same speech stimuli, conditions, and participant populations can generate starkly different patterns of results when different paradigms are used to measure listening effort (Brown & Strand, 2019), suggesting that results obtained from one listening effort task do not necessarily extend to other listening effort tasks. We argue that these findings stem at least in part from weak validity of listening effort measures.

Validity in Listening Effort Research

To gain insights about a theoretical construct, researchers must take steps to ensure that the conclusions drawn from individual studies are valid. In this section, we highlight four forms of validity—internal, external, statistical conclusion, and construct validity—and explain how they are addressed in the listening effort literature. Although all types of validity are important in any research area, we suggest that insufficient attention to construct validity in particular has obscured our understanding of the construct of listening effort.

Different research subdisciplines face different challenges in ensuring the validity of their measures. For example, in some areas of research, there may be nonrandom attrition resulting in systematic differences between groups, participants may guess the experimental hypothesis and act differently as a result, or naturally occurring changes over time may be mistaken for changes due to treatment (see Shadish, Cook, & Campbell, 2002). All of these scenarios threaten internal validity—the extent to which it is possible to make causal inferences about the relationships between variables. Many of these concerns are less relevant in listening effort research, in which studies are often short and occur in a single session (so attrition is rare), the order of conditions can be easily counterbalanced (to avoid order effects), and word lists in different conditions can be matched on relevant lexical variables (or even used in multiple conditions across participants). However, internal validity may still be a concern in some regards; for example, self-report measures of listening effort may be particularly susceptible to demand characteristics (i.e., participants may report increased listening effort in greater levels of background noise because they expect that this is the researcher's intention).

External validity describes the generalizability of findings to different populations or contexts. Listening effort researchers typically appear to be quite cognizant of the fact that results from a particular study may not generalize across different populations (e.g., listening effort may be different depending on age and hearing ability), across different listening conditions (e.g., whether an effect emerges may depend on the difficulty of the task), or across different stimulus materials (e.g., identification of isolated words poses different challenges than connected speech). Further, the presence of a substantial body of individual differences work suggests that researchers often consider the extent to which findings in a particular population are likely to generalize to participants with different perceptual and cognitive abilities.

Statistical conclusion validity describes the extent to which conclusions based on statistical analyses are sound. Much has been written about threats to statistical conclusion validity throughout psychology (e.g., Brysbaert, 2019; McClelland, Lynch, Irwin, Spiller, & Fitzsimons, 2015; Simmons, Nelson, & Simonsohn, 2011), and these concerns certainly apply to the listening effort literature. Indeed, sample sizes are rarely justified, analytical techniques known to reduce statistical power (e.g., median splits; Liben-Nowell, Strand, Sharp, Wexler, & Woods, 2019) are commonplace, and published work often includes many more analyses than would be necessary to address the main hypotheses without reference to whether these decisions were made a priori or whether there were additional undisclosed analyses. However, we would argue that listening effort research is not unique in its tendency to “brush over” statistical conclusion validity, and the recommendations made to improve statistical conclusion validity in other realms (see, for example, García-Pérez, 2012) can easily be applied to the listening effort literature. Therefore, this paper focuses on what we see as the most substantial and unaddressed threat to our understanding of listening effort: construct validity.

Construct validity describes the extent to which an instrument is actually measuring what it purports to measure. A task that is intended to evaluate sustained attention during a listening task but inadvertently measures hearing ability does not give insight into sustained attention, even if the study is thoughtfully designed and its analyses are highly powered. Given the large number and varied nature of the paradigms used to measure listening effort, thorough construct validation is required to ensure that the measures are, in fact, tapping into the same construct.

Convergent Validity & Jingle

Convergent validity is a subtype of construct validity that describes the extent to which multiple instruments intended to measure the same construct are related to one another. There is now ample evidence for very weak convergent validity among measures of listening effort. For example, Johnson et al. (2015) showed that recall and self-report measures of listening effort were uncorrelated ($r = -.21$ – $-.14$). Strand et al. (2018) tested seven measures of listening effort and the average correlation among the measures was $r = .22$, indicating a weak relationship among the tasks (see also Alhanbali et al., 2019). These findings suggest that multiple tasks assumed to assess listening effort may actually be measuring different underlying processes. Edward Thorndike (1904) originally referred to this as the *jingle fallacy*—falsely assuming two instruments measure the same construct because the instruments share a name (i.e., these are all “measures of listening effort”). Indeed, a recent paper on this topic concluded that listening effort measures “should not be used interchangeably as each of them appears to tap into an independent aspect of listening demands” (Alhanbali et al., 2019, p. 13).

The presence of jingle in the literature can lead to inconsistent results across studies that implement different paradigms. If two studies using different instruments reach opposing conclusions, it is unclear whether those findings are actually at odds with each other—indicating a failure to conceptually replicate—or whether they are simply measuring different latent constructs. An analysis of all the conflicting findings in the listening effort literature that may be attributable to weak convergent validity is beyond

the scope of this paper, but one salient example is inconsistent results about the effect of visual cues (i.e., seeing as well as hearing the talker) on listening effort.

There is evidence in the literature that seeing a talker while listening to speech increases listening effort (Alsus et al., 2007; Fraser et al., 2010; Gosselin & Gagné, 2011), decreases listening effort (Mishra et al., 2013b; Rudner, Mishra, Stenfelt, Lunner, & Rönnerberg, 2016; Sommers & Phelps, 2016), has no effect on listening effort (Picou et al., 2011), and has different effects depending on the nature of the noise (Brown & Strand, 2019; Mishra et al., 2013b). It is important to note that conflicting findings in the literature are not necessarily the result of jingle; in addition to differing in their operationalizations of listening effort, these studies also differed in the participant populations, nature of the speech materials, and types of masking noise. However, evidence that the conflicting findings may be due at least in part to task selection comes from the fact that most studies showing that visual cues increase listening effort use dual-task measures (Brown & Strand, 2019; Fraser et al., 2010; Gosselin & Gagné, 2011), whereas most studies showing that visual cues decrease listening effort use recall-based tasks (Rudner et al., 2016; Sommers & Phelps, 2016). Indeed, even in a single population and using the same stimuli and highly-powered analyses, Brown and Strand (2019) showed differing patterns of results when using different tasks: seeing the talker appeared to increase effort when measured with a dual-task paradigm, but had no effect on effort when measured using a recall paradigm. These contradictory findings highlight how different paradigms assumed to measure listening effort may actually be tapping into different aspects of listening effort and/or related processes.

Two well-designed and soundly executed studies that intend to answer the same question may reach opposing conclusions. This may happen simply by chance or because there are meaningful differences in some features like stimuli or participants (i.e., there are hidden moderator variables). However, conflicting results may also occur because although the studies sought to measure the same construct, their choice of instruments led them to measure different underlying constructs (see, for example, Dang, King, & Inzlicht, 2020). In that case, the studies have not actually reached opposing conclusions, they have simply studied different phenomena. *Jingle* in the literature makes it difficult to distinguish between these possibilities. This is especially relevant for conceptual replications: when an original study and a replication use different instruments that are assumed to measure the same underlying construct, a failure to replicate need not imply that one study had a false positive result, but may instead indicate that the instruments do not actually measure the same construct.

Discriminant Validity & Jangle

In addition to convergent validity, another key consideration in establishing construct validity is discriminant validity—the extent to which measurements that are intended to tap into distinct constructs are actually unrelated. For example, performance on a test that is intended to measure listening effort should not instead be primarily measuring fatigue or working memory capacity. However, there is considerable overlap in some tasks intended to measure listening effort and those intended to measure the affective reaction to the increased processing demands generated by difficult listening situations

(see Francis & Love, 2019; Francis & Oliver, 2018 for reviews), as well as those measuring the fatigue that results from expending additional effort.

When tasks measure constructs other than the intended ones, it can lead to the *jangle fallacy*—falsely assuming two instruments must measure different constructs because the instruments have different names (Flake & Fried, 2020; Kelley, 1927). Jangle in the literature can lead to difficulty building and refining theories of listening effort. For example, as mentioned above, the Ease of Language Understanding model outlines a relationship between listening effort and working memory such that in demanding listening situations, listeners with larger working memory capacities are better able to resolve mismatches between the acoustic input and representations in memory, and therefore experience less effort than listeners with smaller working memory capacities (Rönnberg et al., 2013, 2008). Some research has provided support for this claim. For example, Strand et al. (2018) found that performance on the reading span task, a measure of working memory capacity, is correlated with performance on tasks that have been used to measure listening effort including listening span (Sarampalis et al., 2009) and running memory (Sommers & Phelps, 2016) tasks. Additionally, people with larger working memory capacity as measured by a letter monitoring task tend to subjectively report lower levels of listening effort (Rudner, Lunner, Behrens, Thorén, & Rönnberg, 2012).

However, pinpointing whether and how working memory moderates listening effort is complicated by similarity in the tasks used to measure the two constructs. For example, many recall-based listening effort tasks require participants to store and then recall items (e.g., digits, words) presented in noise, just as working memory tasks do. In fact, the listening span task—in which participants listen to sentences, make judgments about them, and then recall the final words of each sentence when prompted—has been used to measure both listening effort (e.g., Johnson et al., 2015; Pichora-Fuller et al., 1995; Sarampalis et al., 2009; Strand et al., 2018) and working memory (e.g., Daneman & Carpenter, 1980). It should not be at all surprising to find a correlation between working memory and listening effort if they are assessed using the same task! In contrast, a study that measured listening effort using a dual-task paradigm (rather than a recall-based one) found no relationship between listening effort and working memory capacity (Brown & Strand, 2018). Thus, the reported relationship between listening effort and working memory capacity may instead be an instantiation of jangle: the two constructs appear to be related because they are often assessed using similar or even identical paradigms.

Questions to Promote Transparency of Measurement Practices (Flake & Fried, 2020)

There has been growing concern about measurement issues in the listening effort literature (e.g., Alhanbali et al., 2019; Strand et al., 2018), as evidenced by papers with titles including: “Listening effort and fatigue: What exactly are we measuring?” (McGarrigle et al., 2014) and “Listening effort: Are we measuring cognition or affect, or both?” (Francis & Love, 2019). Here, we provide recommendations for future work by drawing on transparent psychometric and measurement practices from outside the listening effort literature. A recent paper by Flake and Fried (2020) provides a set of six questions for researchers to consider to promote measurement transparency. Below, we discuss how these questions apply to the listening effort literature and make suggestions

for how listening effort researchers can improve measurement practices. Addressing these questions will not only enable other researchers to evaluate and build on previous work, but will also facilitate future meta-analyses to provide a clearer picture of the literature without being clouded by questionable measurement practices (Flake & Fried, 2020).

What Is Your Construct?

Before a researcher can measure a construct, they must identify conceptually what they intend to measure. This conceptual definition is likely to be informed by the theoretical framework and underlying assumptions about the construct. Conceptual definitions must precede measurement decisions, as a clear understanding of the construct is essential for designing a paradigm that captures that intended construct. For example, Herrmann and Johnsrude (2020) distinguish between “effort” (a subjective experience) and “engagement” (the act of investing resources in an activity). According to these definitions, at low levels of task difficulty it is possible to engage resources without experiencing subjective effort. Defining listening effort as a subjective experience therefore warrants a different type of measurement tool than defining effort according to the recruitment of cognitive resources.

Listening effort has been defined in multiple ways, including:

- “the attentional requirements necessary to understand speech” (Bourland-Hicks & Tharpe, 2002; Fraser et al., 2010)
- “the cognitive resources allocated for speech recognition” (Picou et al., 2011)
- “the mental exertion required to attend to, and understand, an auditory message ” (McGarrigle et al., 2014)
- “the deliberate allocation of mental resources to overcome obstacles in goal pursuit when carrying out a [listening] task” (Pichora-Fuller et al., 2016)
- “the amount of processing resources (perceptual, attentional, cognitive, etc.) allocated to a specific auditory task, when the task demands are high (adverse listening conditions) and when the listener strives to reach a high-level of performance on the listening task” (Gagné et al., 2017a)

Although these definitions share many features, the differences among them reveal unresolved theoretical questions. Does listening effort depend solely on attentional resources or on other cognitive resources (e.g., working memory, executive control) as well? Can listening effort occur only when a listener strives for high levels of performance? Do listeners expend any effort in pristine listening conditions, or is effort only required in adverse listening conditions? Does effort arise from merely attending to an auditory message (e.g., early segregation of a relevant auditory stream) or from understanding the message (e.g., downstream comprehension), or both? Is listening effort speech-specific or can it occur when listening to other auditory signals (e.g., music)? Is effort an experience or the mental act of recruiting resources (see Herrmann & Johnsrude, 2020)? Thus, the first challenge researchers face is making a decision about which of the many definitions best describes their intended construct.

Further complicating defining the construct is that listening effort may be difficult to distinguish from other related constructs such as “mental effort” (Panico & Healey, 2009), “perceptual effort” (Tun et al., 2009), “cognitive effort” (Obleser, Wöstmann, Hellbernd, Wilsch, & Maess, 2012; Piquado, Isaacowitz, & Wingfield, 2010), “cognitive load” (Zekveld et al., 2011), and “processing load” (Kramer et al., 2012; Zekveld et al., 2010). It may be that the different terminology used across studies simply reflects differences in word choice, and these terms are actually being treated synonymously with listening effort. Alternatively, researchers may have explicit reasons for using one term rather than another; for example, a study may use a listening task to induce effort but also wish to draw connections to effortful tasks that do not require listening. Therefore, the researchers may want to describe the research as assessing more general “cognitive effort” rather than “listening effort.” In any case, we recommend that researchers provide a clear justification for why they chose the particular term and definition they did.

Further, given the evidence that dual-task, physiological, and self-report measures of listening effort may not be measuring the same underlying construct (Alhanbali et al., 2019; McGarrigle et al., 2014), researchers should consider being more specific about what is being measured by alluding to the task when they define the construct. That is, rather than assuming that a particular task is a measure of listening effort, researchers may choose to specify that dual-task paradigms assess the dual-task costs of listening, physiological paradigms reflect a physiological correlate of a change in listening demand, and subjective measures indicate perceived listening effort (McGarrigle et al., 2014). These more specific constructs may be subcomponents of listening effort (just as working memory and recognition memory are subcomponents of memory), and it is likely useful to consider them as such. Regardless, explicitly defining the construct of interest will help build consistency in the literature.

Why and How Do You Select Your Measure?

After defining the construct, the researcher must make the critical decision of how to measure it. As we have noted, this is not a simple task when there are many choices of measures and no agreed “gold standard.” Indeed, researchers rarely justify the listening effort measure they use, thereby making it difficult for future researchers to know what criteria to use when selecting a measure for their own study. Some factors that researchers might consider when justifying the use of a listening effort measure include the settings or applications of interest, consistency with prior work, validity evidence for the measure (in cases where it is available), and simply access to materials.

Regarding setting or application of interest, the choice of listening effort measure may differ between a study focused on clinical outcomes compared to a study concerned with educational settings. In clinical work, researchers may be more likely to use a self-report measure, as they are likely more interested in patients’ subjective experiences of effort than how effortful listening affects recall of what was heard. In an educational setting, however, it may be more appropriate to use a behavioral measure of listening effort that tests retention of information, as that more closely approximates how listening effort can impact learning. Further, many dual-task measures assess instantaneous listening effort, whereas recall measures typically capture how listening effort impacts memory later in time, and one of these time frames may be of greater interest to the researcher.

In other situations, researchers may choose a particular listening effort measure in order to be consistent with prior work. For example, if a researcher is interested in the extent to which an effect generalizes to other listening conditions or participants, they may choose to use the same measure that was used in the previous work to make the results most comparable. Still in other cases, researchers may justify their use of a measure based on what equipment they have access to. For instance, if a study called for the use of a physiological measure and the researchers did not have access to EEG technology, they may elect to use a different physiological measure such as pupillometry. There are many reasons a researcher may choose a particular measure, but regardless of the reason, it is helpful to readers and future researchers to know the rationale behind the decision.

Flake and Fried (2020) also recommend ensuring that there is alignment between the selected measure and the theoretical definition proposed. This allows researchers to be more explicit about what is causing changes in the outcome. For example, a seminal study by Rabbitt (1968) argued that noise disrupts the digit rehearsal process; however, subsequent interpretations of this finding have been inconsistent. These impairments in recall are sometimes attributed to disrupted rehearsal (McCoy et al., 2005; Pichora-Fuller et al., 1995), encoding (Brown & Strand, 2019; Sommers & Phelps, 2016; Sommers, Tye-Murray, Barcroft, & Spehar, 2015), or both (Murphy, Craik, Li, & Schneider, 2000). Critically, all of these papers cite the same Rabbitt (1968) study as support for their reasoning. Although these differences in interpretation may seem minor, our understanding of how background noise affects recall has implications for which recall-based listening effort tasks we use. For example, if difficult listening conditions interfere primarily with *rehearsal*, as Rabbitt (1968) suggested, then recall-based tasks that rely on successful rehearsal should be impaired (e.g., serial recall tasks, running memory tasks). In contrast, if difficult listening conditions interfere with the process of *encoding*, then only recall-based tasks that require encoding information into long-term memory (e.g., retention for information from long passages)—and *not* those that only require short-term rehearsal of information (e.g., serial recall tasks using short lists of digits)—would be expected to show effects of listening effort. Thus, the reason for selecting a particular measure may be driven by the extent to which the paradigm taxes the processes (e.g., encoding or rehearsal) that are assumed to be affected by changes in listening effort.

Finally, Flake and Fried (2020) note that researchers should report any information about the reliability or validity of the measure from prior work—internal consistency, evidence for convergent validity, etc.—and describe why they think the evidence should extend to the particular context of their experiment. These recommendations certainly apply to the listening effort literature—there are only vague traces of validity evidence in the literature, and it is commonplace for the same instrument to be used in a variety of contexts. For example, the same dual-task paradigm has been used to assess changes in listening effort associated with noise-reduction algorithms in hearing aids (Sarampalis et al., 2009), changes in background noise level (albeit with different types of background noise; Brown & Strand, 2018; Picou & Ricketts, 2014), and differences in talker characteristics (Brown, McLaughlin, Strand, & Van Engen, 2020). If this paradigm measures the intended construct of listening effort, and those manipulations affect the amount of effort listeners expend to understand speech, then it is perfectly reasonable to expect the

validity evidence to generalize across these contexts. However, this should be explicitly stated and the rationale justified (Flake & Fried, 2020).

What Measures Did You Use?

Once the measure is selected, Flake and Fried (2020) recommend transparently detailing every aspect of its implementation. Although clearly reporting methodological detail is important in all research, it is particularly important for research on listening effort given the large number of paradigms that are regularly implemented, as well as variability in administration procedures, stimulus selection, and analysis techniques. Researchers should report where the measure came from and details of the paradigm's implementation, including the particular stimuli used, acoustic characteristics of the talker, and the listening conditions under which participants were tested. Even for two studies that appear on the surface to use the same paradigm, details of the implementation of the paradigm often differ. Therefore, we recommend explicitly reporting the following details:

- How was the speech presented to participants? Were headphones used? What kind?
- How were the speech files edited? Was noise removed? Were the files leveled? How? What software was used?
- What were the characteristics of the talker?
- What was the response format? Were responses to speech stimuli reported verbally, typed, or indicated via button press or some other mechanism?
- What was the nature of the stimuli? Were the speech stimuli syllables, words, or sentences? Were the sentences semantically constraining? How were words chosen?
- What were the listening conditions like? What kind of background noise was used? How was the noise created? Was testing conducted in a sound attenuating room?
- How were participants given instructions? Were they instructed to guess when unsure?
- What was the interstimulus interval between items?
- What was presented on the screen during the listening task?
- Precisely how many items did each participant respond to in each condition, and how many observations were included in the final analysis?

The most comprehensive way to ensure the methods are fully transparent is to make all materials publicly available in an online repository such as the Open Science Framework. There may be methodological details that seem trivial when writing a method section, but at a later stage are revealed to be influential. For example, an experimenter may select lists of stimuli that are matched on several lexical variables, but future researchers may realize that the lists differed in a lexical variable that was not considered, thereby threatening internal validity. Sharing materials (including speech stimuli and programs for stimulus presentation) enables any notable differences between studies to be identified and helps to ensure that future replication attempts are appropriately conducted. Of course, researchers should scrutinize materials obtained from other sources as closely as they would their own to ensure that they are not blindly reusing faulty materials and perpetuating unnoticed confounds.

How Do You Quantify Your Measure?

After a task has been selected, researchers must make decisions about how the dependent variable is quantified. Researchers should report precisely what the dependent variable is, describe and justify any transformations that were made to the data prior to analysis (e.g., log-transformed response times, standardized outcomes, rationalized arcsine transformations, etc.), and state and justify any aggregation that occurred to arrive at numbers that were included in the final analyses. It is also necessary to clearly define the criteria for identifying and removing outliers both at the level of the participant and the level of the trial. Of the response time dual-task measures of listening effort cited in a recent review of the literature (Gagné et al., 2017a), 88% did not report whether any outliers were excluded from analyses or describe exclusion criteria. Although these details may seem minor, making different decisions about which data-points to exclude is a *researcher degree of freedom* that may affect outcomes (Simmons et al., 2011).

As one example, although the running memory task is commonly used to measure listening effort, the outcome is quantified differently across studies. In this task, participants are aurally presented lists of words for later recall, but given that the speech is often presented in background noise, participants may not correctly identify some words and therefore cannot recall those items. A decision must therefore be made about how to account for incorrect identification to ensure that recall performance is not confounded with speech identification performance. When the speech is presented in noise, researchers typically give participants credit for recalling a word as it was perceived (e.g., Brown & Strand, 2019; Johnson et al., 2015; Ng et al., 2013; Pichora-Fuller et al., 1995; Sarampalis et al., 2009) or adjust for recall performance of the last item in the list (which should be perfect if the word was correctly identified; Strand et al., 2018). When speech is presented in quiet, some studies make no adjustments (Sommers & Phelps, 2016) and still others only analyze lists in which the last word was correctly identified (McCoy et al., 2005). These are all reasonable choices, but these decisions should be explicitly noted and justified. Finally, we recommend that researchers post their raw data and analysis code in a repository such as the Open Science Framework whenever possible. As with sharing materials, this will help to increase the transparency and reproducibility of our research by allowing future researchers to thoroughly understand the method by which the data were processed.

Do You Modify the Measure? If So, How and Why?

In addition to reporting an experiment's methodology in sufficient detail such that another researcher could replicate the study, researchers should explicitly note any modifications to the measure and describe why these modifications were necessary. For example, in one dual-task measure of listening effort, participants listen to and respond to speech as they simultaneously classify visually-presented numbers as even or odd via button press, and response times to the number task are taken as an indication of effort. However, in some versions of this task the speech and numbers are presented asynchronously so they do not always coincide (Brown & Strand, 2018; Sarampalis et al., 2009), and in others experimental constraints ensure that number trials only occur during presentation of the word or sentence (Brown et al., 2020; Picou & Ricketts, 2014). Two tasks are more likely to draw from the same pool of cognitive resources (as is assumed in

the dual-task paradigm) if they are performed simultaneously, so differences across studies may be attributable to the timing of the two tasks.

Further, researchers may opt to use a different number of critical items or probe trials (cf., Brown & Strand, 2018; Picou & Ricketts, 2014), and this may affect outcomes because experiments that are longer or more demanding may lead to more effort. As one final example, studies that employ recall measures of listening effort often differ in the extent to which participants have prior knowledge of how many items will be presented on a given trial, with some studies using a fixed number of items (Johnson et al., 2015; Sarampalis et al., 2009) and others allowing list length to vary (Pichora-Fuller et al., 1995; Strand et al., 2018; Zekveld & Kramer, 2014), which may affect the rehearsal and recall strategies adopted by participants during the task. There is not clear evidence that any of these choices are better or worse than others, but these modifications should be explicitly noted and justified to facilitate comparing the results of studies using modified versions of the same task.

Do You Create the Measure on the Fly?

Finally, researchers should indicate whether the measure was created for that particular experiment and, if so, justify the decision to use a new measure in lieu of an existing one. Ideally, if a new measure is used the researchers should validate it in a preliminary experiment such as testing it against an existing measure to assess convergent validity (see “Additional Recommendations” below). Alternatively, researchers may first test the measure in a familiar context in which the pattern of results can be anticipated based on previous research before applying it to a novel situation (i.e., include a positive control). For example, before using a vibrotactile dual-task measure to assess how listening effort differed in audio-only versus audiovisual conditions, a recent study from our lab (Brown & Strand, 2019) first tested the vibrotactile task in an audio-only condition with multiple levels of background noise. This enabled us to confirm that the task was sensitive to changes in noise—which have been robustly demonstrated to affect listening effort—before testing it in a novel context. If these approaches are not possible, any existing validity evidence should be discussed and a lack of validity evidence should be noted as a limitation of the study.

Creating measures on the fly (e.g., new subjective measures, new secondary tasks in dual-task studies, etc.) is commonplace in the listening effort literature, and may be necessary for the purposes of a particular study, but these decisions are rarely justified and the measures are rarely validated. Thus, the listening effort literature would benefit greatly from transparency about where particular measures came from, and if they are brand new, supporting validation evidence (or lack thereof) should be described.

Additional Recommendations

In addition to the recommendations laid out by Flake and Fried (2020) that we have applied to work on listening effort, our review of the literature has highlighted several other issues that when addressed will strengthen listening effort research. First, we recommend looking for ways to assess convergent validity within the context of a particular study. Papers that have opted to include multiple measures of listening effort in the same study have been very informative in pointing out the presence of jingle in the

literature; indeed, correlations among measures of listening effort tend to be small, suggesting that the various measures are not tapping into the same underlying construct (e.g., Alhanbali et al., 2019; Johnson et al., 2015; Seeman & Sims, 2015). Thus, we recommend that whenever possible researchers include multiple measures of listening effort in a given study, even if measurement issues are not the focus of the study. This may help to identify issues pertaining to convergent validity, resolve inconsistencies in the literature, and help clarify what various measures of listening effort are in fact assessing. Even when it is not feasible to add substantially to the study design, including a subjective measure asking participants to self-report levels of effort will provide insight into whether the experimental manipulation elicited changes in the subjective experience of effort.

To researchers who may be reluctant to add additional components to their studies, we would point out that studies including multiple measures have the ability to provide methodological insights in addition to shedding light on the theoretical question of interest, and are therefore more likely to be published than a study that reports null findings using a single measure without supporting validity evidence. However, it is worth noting that including multiple measures increases the number of possible analyses and therefore the likelihood of producing spurious findings. This (and other forms of analytic flexibility) can be addressed by pre-registering analysis plans (Hales, Wesselmann, & Hilgard, 2018) to make it clear which measures and analyses are of primary interest and, when appropriate, statistically control for multiple comparisons.

Our second recommendation is that researchers consider assessing discriminant validity directly. One method of doing this is to include additional measures that are not intended to capture listening effort—such as working memory tasks—to demonstrate that they do not correlate with the listening effort measure. As another example, studies that assess the extent to which changes in the level of the background noise affect listening effort tend to hold the level of the speech constant while varying the level of the background noise. It is conceivable, then, that poorer performance on the task assumed to measure listening effort (e.g., a secondary task) does not actually reflect increases in listening effort, but instead reflects noise-induced performance declines on cognitive tasks more generally. Thus, it is important to establish that performance on the secondary task (e.g., making judgments about whether visually-presented numbers on a screen are odd or even) changes as a function of background noise when speech is present but not when it is absent. We recently showed that performance on a commonly used dual-task paradigm differed predictably when speech was presented in an easy and a difficult level of background noise, but was unaffected by background noise in the absence of speech (Brown & Strand, 2018). That is, categorizing visually-presented numbers as odd versus even was not impaired by adding background noise, unless participants also had to listen to speech presented in the noise (i.e., perform a simultaneous listening task). We recommend that researchers conduct these sorts of tests before conducting a listening effort experiment (or cite other papers that have obtained this validity evidence already), because if performance on the task is affected by background noise to a similar extent regardless of whether speech is present, then that measure is likely not assessing the intended construct. Addressing jangle in this way will help to clarify what tasks are actually measuring and may help refine theory.

Relatedly, it is important that researchers consider the possibility that even if their measurement tool is indeed assessing the intended construct, it may be simultaneously assessing another construct as well. As one example, pupillometry has been used to measure both listening effort over time (reduced pupil dilation over the course of the experiment is assumed to reflect reduced listening effort; Brown et al., 2020; Porretta & Tucker, 2019) and fatigue over time (reduced pupil dilation is assumed to reflect increased fatigue; McGarrigle, Dawes, Stewart, Kuchinsky, & Munro, 2016). As a measure of changes in physiological arousal patterns, pupillometry is likely sensitive to both mental effort exertion and the experience of fatigue (along with other factors, see Zekveld et al. 2018 for a review), and inferring which of the constructs is at play may depend on the nature of the task and stimuli. Researchers should therefore carefully consider whether a particular measurement tool may be sensitive to multiple constructs, and ensure that the constructs can be delineated in the context of a particular experiment (see McGarrigle et al., 2016 for more on this issue).

Third, we believe that research on listening effort (and many other constructs) would benefit from regular use of Constraints on Generality statements in manuscripts (Simons, Shoda, & Lindsay, 2017). In these sections, researchers specify the target populations and particular contexts to which they expect their findings to generalize, including statements about participants, materials, and procedures that are deemed critical. Thus, Constraints on Generality sections may be useful in encouraging authors and readers to explicitly consider the external validity of the studies. Simons et al. (2017) argue that Constraints on Generality sections help to: 1) protect authors from making overly bold claims that cloud the literature with findings that claim to be more generalizable than they actually are, 2) make it more likely that findings will replicate, given that the boundary conditions are specified, and 3) inspire follow-up studies that test the specified limits of the effect. In the listening effort literature, it would be particularly useful to have Constraints on Generality sections in which researchers specify whether they have reason to believe that their findings would emerge with other measures (thereby helping to identify assumptions about the constructs that particular tasks are measuring), or would apply to other populations, stimuli, noise conditions, and so forth.

Fourth, we recommend that when writing literature reviews, authors pay careful attention to the measures that are used in the studies they are citing and, when appropriate, explicitly mention the measures used in the studies being reviewed. In addition to providing other researchers with the level of detail they need to make informed decisions about measurement selection for their own experiments, this practice will help others identify patterns that may account for discrepancies in the literature. For example, as we described above, the literature appears to be quite mixed regarding the influence of audiovisual relative to audio-only speech on listening effort, but the discrepancies may simply be attributable to different operationalizations of listening effort across studies. Thus, when drawing conclusions across studies, it is important that researchers do not assume the absence of jingle.

Although the focus of this paper is on ways to strengthen validity in listening effort research, our final recommendation is that researchers consider the reliability of their measurement tools before conducting an experiment, and report that reliability in published work. A recent paper noted that behavioral measures of cognitive processing rarely receive the psychometric scrutiny that other areas within the field of psychology—

particularly those that rely heavily on surveys—receive (Parsons, Kruijt, & Fox, 2019). The authors argue that cognitive psychologists should routinely provide measures of reliability, and we agree with this stance; although validity and reliability are distinct psychometric properties, reliability places a limit on validity. In other words, if a measure has poor reliability, it cannot have strong validity (i.e., if an outcome correlates weakly with itself, it cannot possibly correlate strongly with other outcomes). Despite its fundamental role in psychological measurement, reliability is rarely acknowledged in listening effort research. Luckily, given that this research area typically involves repeated measurements within participants, linear mixed effects models provide a straightforward solution to this shortcoming: the intraclass correlation—a reliability estimate that can be easily extracted from these models (see Brysbaert, 2019 for a tutorial). Given the push away from ANOVAs and toward mixed effects models in speech research, reporting intraclass correlation coefficients would be a relatively easy way to provide evidence about reliability in published work.

Conclusions

At the time of writing, a Google Scholar search for the term “listening effort” rendered almost 6,000 results. Despite the large number of published studies about listening effort, research on the construct has not generally placed much emphasis on establishing validity evidence for our measures. Poor measurement practices hinder our ability to interpret the results of individual studies as well as weaken the foundation upon which theories are built. The current state of the literature need not imply that listening effort is not a “real” construct; indeed, most people have experienced the feelings of strain and effort associated with listening in adverse conditions, as well as difficulty multitasking and recalling what was heard in such situations. We argue instead that the inconsistent findings may reflect the fact that the tasks used to measure listening effort are not tapping into the same underlying construct, and a lack of validation work on listening effort measures has concealed this fact. We encourage the community of listening effort researchers to place greater emphasis on measurement issues in papers they write, review, and edit to ensure that the next 6,000 articles we collectively publish demonstrate the methodological rigor that is necessary to form consensus in the literature.

Acknowledgments

Portions of this work were presented at the Auditory Perception, Cognition, and Action Meeting (2019). We are grateful to Ronan McGarrigle for helpful feedback on a previous version of this paper.

Disclosure Statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Carleton College, a National Institutes of Health grant via the National Institute on Deafness and Communication Disorders awarded to Julia Strand (R15-

DC018114), and the National Science Foundation through a Graduate Research Fellowship awarded to Violet Brown (DGE-1745038)

ORCID

Julia F. Strand  <http://orcid.org/0000-0001-5950-0139>

Naseem H. Dillman-Hasso  <http://orcid.org/0000-0002-8284-4383>

Violet A. Brown  <http://orcid.org/0000-0001-5310-6499>

References

- Alhanbali, S., Dawes, P., Millman, R. E., & Munro, K. J. (2019). Measures of listening effort are multidimensional. *Ear and Hearing*. doi:10.1097/AUD.0000000000000697
- Alsius, A., Navarra, J., & Soto-Faraco, S. (2007). Attention to touch weakens audiovisual speech integration. *Experimental Brain Research*, 183(3), 399–404.
- Barry, A. E., Chaney, B., Piazza-Gardner, A. K., & Chavarria, E. A. (2014). Validity and reliability reporting practices in the field of health education and behavior: A review of seven journals. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, 41(1), 12–18.
- 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: JSLHR*, 45(3), 573–584.
- Broadbent, D. E. (1958). The effects of noise on behavior. In D. E. Broadbent (Ed.), *Perception and communication* (pp. 81–107). Elmsford, NY, US: Pergamon Press.
- Brown, V. A., McLaughlin, D. J., Strand, J. F., & Van Engen, K. J. (2020). Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort. *The Quarterly Journal of Experimental Psychology*. doi:10.1177/1747021820916726
- Brown, V. A., & Strand, J. F. (2018). Noise increases listening effort in normal-hearing young adults, regardless of working memory capacity. *Language, Cognition and Neuroscience*, 34, 628–640.
- Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition*, 2(1). doi:10.5334/joc.89
- Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, 2(1), 187.
- Cohen, R. J., & Swedlik, M. (2010). *Psychological testing and assessment: An introduction to tests and measurement*. New York: McGraw Hill.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281–302.
- Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19, 450–466.
- Dang, J., King, K. M., & Inzlicht, M. (2020). Why are self-report and behavioral measures weakly correlated? *Trends in Cognitive Sciences*, 24(4), 267–269.
- 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.
- Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*. doi:10.31234/osf.io/hs7wm
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370–378.
- Francis, A. L., & Love, J. (2019). Listening effort: Are we measuring cognition or affect, or both? *Wiley Interdisciplinary Reviews. Cognitive Science*, 11(1), e1514.

- Francis, A. L., & Oliver, J. (2018). Psychophysiological measurement of affective responses during speech perception. *Hearing Research*, 369, 103–119.
- Fraser, S., Gagné, J.-P., Alepins, M., & Dubois, P. (2010). Evaluating the effort expended to understand speech in noise using a dual-task paradigm: The effects of providing visual speech cues. *Journal of Speech, Language, and Hearing Research: JSLHR*, 53(1), 18–33.
- Gagné, J.-P., Besser, J., & Lemke, U. (2017a). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, 21, 2331216516687287.
- Gagné, J.-P., Besser, J., & Lemke, U. (2017b). Behavioral assessment of listening effort using a dual-task paradigm: A review. *Trends in Hearing*, 21, 2331216516687287.
- García-Pérez, M. A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, 3, 325.
- Gosselin, P. A., & Gagné, J.-P. (2011). Older adults expend more listening effort than young adults recognizing audiovisual speech in noise. *International Journal of Audiology*, 50(11), 786–792.
- Hales, A. H., Wesselmann, E. D., & Hilgard, J. (2018). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*. doi:10.1007/s40614-018-00186-8
- Herrmann, B., & Johnsrude, I. S. (2020). A model of listening engagement (MoLE). *Hearing Research*, 397, 108016.
- 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.
- Kelley, T. L. (1927). *Interpretation of educational measurements*. 353. <https://psycnet.apa.org/fulltext/1928-00533-000.pdf>
- Kramer, S. E., Lorens, A., Coninx, F., Zekveld, A. A., Piotrowska, A., & Skarzynski, H. (2012). Processing load during listening: The influence of task characteristics on the pupil response. *Language and Cognitive Processes*, 28(4), 426–442.
- 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.
- Liben-Nowell, D., Strand, J., Sharp, A., Wexler, T., & Woods, K. (2019). The danger of testing by selecting controlled subsets, with applications to spoken-word recognition. *Journal of Cognition*, 2, 1.
- 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.
- McClelland, G. H., Lynch, J. G., Irwin, J. R., Spiller, S. A., & Fitzsimons, G. J. (2015). Median splits, Type II errors, and false-positive consumer psychology: Don't fight the power. *Journal of Consumer Psychology: The Official Journal of the Society for Consumer Psychology*, 25(4), 679–689.
- McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., & Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *The Quarterly Journal of Experimental Psychology. A, Human Experimental Psychology*, 58(1), 22–33.
- 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.
- Mishra, S., Lunner, T., Stenfelt, S., Rönnerberg, J., & Rudner, M. (2013a). Visual information can hinder working memory processing of speech. *Journal of Speech, Language, and Hearing Research*, 56, 1120–1132.

- Mishra, S., Lunner, T., Stenfelt, S., Rönnerberg, J., & Rudner, M. (2013b). Seeing the talker's face supports executive processing of speech in steady state noise. *Frontiers in Systems Neuroscience*, 7, 96.
- 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., & Rönnerberg, 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.
- Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common oscillatory network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 32(36), 12376–12383.
- Pals, C., Sarampalis, A., & Baskent, D. (2013). Listening effort with cochlear implant simulations. *Journal of Speech, Language, and Hearing Research: JSLHR*, 56(4), 1075–1084.
- Panico, J., & Healey, E. C. (2009). Influence of text type, topic familiarity, and stuttering frequency on listener recall, comprehension, and mental effort. *Journal of Speech, Language, and Hearing Research: JSLHR*, 52(2), 534–546.
- Parsons, S., Kruijt, A.-W., & Fox, E. (2019). Psychological science needs a standard practice of reporting the reliability of cognitive-behavioral measurements. *Advances in Methods and Practices in Psychological Science*, 2(4), 378–395.
- Pichora-Fuller, M. K. (2016). How social psychological factors may modulate auditory and cognitive functioning during listening. *Ear and Hearing*, 37(Suppl 1), 92S– 100S.
- 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: JSLHR*, 54(5), 1416–1430.
- Piquado, T., Isaacowitz, D., & Wingfield, A. (2010). Pupillometry as a measure of cognitive effort in younger and older adults. *Psychophysiology*, 47(3), 560–569.
- Porretta, V., & Tucker, B. V. (2019). Eyes wide open: Pupillary response to a foreign accent varying in intelligibility. *Frontiers in Communication*, 4(8).
- Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology*, 20(3), 241–248.
- Rönnerberg, 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.
- Rönnerberg, 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(sup2), S99–S105.
- Rudner, M., Lunner, T., Behrens, T., Thorén, E. S., & Rönnerberg, 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., Mishra, S., Stenfelt, S., Lunner, T., & Rönnerberg, J. (2016). Seeing the talker's face improves free recall of speech for young adults with normal hearing but not older adults with hearing loss. *Journal of Speech, Language, and Hearing Research: JSLHR*, 59(3), 590–599.
- Sarampalis, A., Kalluri, S., Edwards, B., & Hafter, E. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *Journal of Speech, Language, and Hearing Research: JSLHR*, 52(5), 1230–1240.

- Seeman, S., & Sims, R. (2015). Comparison of psychophysiological and dual-task measures of listening effort. *Journal of Speech, Language, and Hearing Research: JSLHR*, 58(6), 1781–1792.
- Shadish, W. R., Jr., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 12(6), 1123–1128.
- Sommers, M. S., & Phelps, D. (2016). Listening effort in younger and older adults: A comparison of auditory-only and auditory-visual presentations. *Ear and Hearing*, 37(Suppl 1), 62S– 8S.
- Sommers, M. S., Tye-Murray, N., Barcroft, J., & Spehar, B. P. (2015). The effects of meaning-based auditory training on behavioral measures of perceptual effort in individuals with impaired hearing. *Seminars in Hearing*, 36(4), 263–272.
- Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11), 1.
- Strand, J. F., Brown, V. A., & Barbour, D. L. (2020). Talking points: A modulating circle increases listening effort without improving speech recognition in young adults. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-020-01713-y
- Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research: JSLHR*, 61, 1463–1486.
- Thorndike, E. L. (1904). *An introduction to the theory of mental and social measurements*. Teacher's College, Columbia University, Science Press, New York.
- 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.
- Zekveld, A. A., Koelewijn, T., & Kramer, S. E. (2018). The Pupil Dilation Response to Auditory Stimuli: Current State of Knowledge. *Trends in Hearing*, 22, 2331216518777174.
- 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.
- 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.