

JANUARY 05 2023

Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations^{a)}

Melissa M. Baese-Berk; Susannah V. Levi ; Kristin J. Van Engen



J Acoust Soc Am 153, 68–76 (2023)

<https://doi.org/10.1121/10.0016806>



View
Online



Export
Citation

CrossMark



**Advance your science and career
as a member of the**

ACOUSTICAL SOCIETY OF AMERICA

LEARN MORE



Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations^{a)}

Melissa M. Baese-Berk,^{1,b)} Susannah V. Levi,²  and Kristin J. Van Engen³

¹Department of Linguistics, University of Oregon, Eugene, Oregon 97403, USA

²Department of Communicative Sciences and Disorders, New York University, New York, New York 10012, USA

³Department of Psychological and Brain Sciences, Washington University in St. Louis, St. Louis, Missouri 63130, USA

ABSTRACT:

Intelligibility measures, which assess the number of words or phonemes a listener correctly transcribes or repeats, are commonly used metrics for speech perception research. While these measures have many benefits for researchers, they also come with a number of limitations. By pointing out the strengths and limitations of this approach, including how it fails to capture aspects of perception such as listening effort, this article argues that the role of intelligibility measures must be reconsidered in fields such as linguistics, communication disorders, and psychology. Recommendations for future work in this area are presented. © 2023 Acoustical Society of America.

<https://doi.org/10.1121/10.0016806>

(Received 30 March 2022; revised 8 December 2022; accepted 18 December 2022; published online 5 January 2023)

[Editor: Benjamin V. Tucker]

Pages: 68–76

I. INTRODUCTION

Although speech communication is often experienced as relatively effortless, a variety of common listening circumstances make this process more challenging. While listeners typically have little difficulty understanding speech from familiar talkers and familiar accents in quiet listening conditions, they demonstrate more difficulty when listening to speech in noise, speech from unfamiliar talkers, or speech from unfamiliar accents.

Evidence for an increased challenge in understanding certain types of speech is typically drawn from intelligibility measures, which assess a listener's ability to transcribe or repeat the speech they hear by measuring the accuracy with which the listener's response matches the target utterance. Intelligibility tasks are viewed as objective measures of a listener's performance, as they provide a binary judgment for whether or not a listener correctly identified the intended target. These measures, thus, characterize listeners' ability to achieve the end point goal of speech perception: correctly identifying the linguistic items (phonemes, words, sentences) produced by the speaker. Of course, this goal of speech perception could be measured in multiple ways. That is, a listener's ability to directly transcribe or repeat the sounds, words, or phrases they hear (i.e., intelligibility) is only part of the process of "understanding" speech. For example, a listener could also be asked to provide objective evidence of how much of the speech they remember or understand via rephrasing the utterance or answering questions about it. Alternatively, they could be asked to subjectively rate the

difficulty in understanding speech (i.e., "comprehensibility"; Munro and Derwing, 1995a). Altogether, speech perception is a multi-faceted task, and intelligibility measures only capture a portion of the phenomenon.

Even so, evidence for listening challenges¹ tends to be rooted in intelligibility measures: listeners' responses typically include fewer correct phonemes, words, or sentences in the more challenging circumstances (e.g., Levi, 2015; Nygaard and Pisoni, 1998). Objective intelligibility measures are also used widely to assess change over time as a function of practice on a task or exposure to talkers or accents (e.g., Bradlow and Bent, 2008)² or to compare various listening situations to one another (e.g., Bent *et al.*, 2016; Borrie *et al.*, 2017; McLaughlin *et al.*, 2018). Further, intelligibility is used as a measure to classify differences in listener groups [e.g., age (Pichora-Fuller *et al.*, 1995) and hearing status (Suter, 1985)]. Taken together, intelligibility measures have often been used as direct measures of speech perception and of the challenges of particular listening circumstances, without substantial critical reflection on the limitations of such a measure.

However, recent work suggests that intelligibility measures capture only part of the challenge listeners face (Brown *et al.*, 2020; McLaughlin and Van Engen, 2020; Winn and Teece, 2021). Further, they fail to accurately assess *what* makes these listening circumstances challenging and do not on their own provide insight into how to alleviate these challenges for listeners.

In this position piece on reconsidering classic ideas in speech communication, we explore what intelligibility, as a measure, has shown us about speech perception and what it fails to capture. We address recent evidence that, even when speech intelligibility measures do not reveal differences in

^{a)}This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.

^{b)}Electronic mail: mbaesebe@uoregon.edu

performance on a task, other measures, such as listening effort, comprehensibility, and working memory, do demonstrate differences, suggesting potential dissociations between these related measures. We conclude with recommendations for future work that will help us better understand the challenges of speech perception and the downstream implications of those challenges.

II. BENEFITS OF INTELLIGIBILITY MEASURES

The appeal of intelligibility as a metric of speech perception is clear. First, it is an objective measure that allows for easy comparisons across individual listeners, across listener groups, across types of speech, and across listening circumstances more generally. Second, it is a simple output measure: did the listener accurately represent the words (or string of phonemes) that were spoken? Third, it provides a measure of the goal of communication—identifying the linguistic content—specifically by quantifying the sounds, words, or phrases a listener heard.

Anyone can tell you that listening to speech in a noisy environment is more challenging than listening to speech in a quiet environment. This intuition is clearly captured by intelligibility measures: speech in noise results in reduced intelligibility compared to speech in quiet [e.g., Cherry, 1953; see McDermott (2009) for a review]. This basic finding has led to a large body of research investigating aspects of speech in noise, including the language of the noise in the background (Calandruccio *et al.*, 2010; Van Engen and Bradlow, 2007), language experience of the talker or listener (Brouwer *et al.*, 2012; Cooke and Garcia Lecumberri, 2018; Mayo *et al.*, 1997; Van Engen, 2010; Van Wijngaarden *et al.*, 2002), type of background noise (Helfer and Freyman, 2014; Van Engen *et al.*, 2014; Vermiglio *et al.*, 2019), and effects of spatial separation of noise and targets (Arbogast *et al.*, 2002; Freyman *et al.*, 2001). Similarly, reverberation also results in reduced intelligibility of the speech signal (Crum, 1974; Nabelek and Pickett, 1974). Each of these findings has been supported through the use of *intelligibility* as the key measure, and this broad investigation has allowed researchers to better understand precisely why speech in noise is difficult for listeners and what factors may impact listening, either positively or negatively.

Intelligibility has also been used to understand challenges resulting from variation in talkers. Speech produced by second-language learners, for example, tends to be less intelligible than speech produced by native speakers (Lane, 1963; Munro and Derwing, 1995a). Similarly, regional, but unfamiliar, varieties of speech tend to result in lower intelligibility than familiar varieties (e.g., Adank *et al.*, 2009). Talkers with speech disorders (e.g., dysarthria) are generally less intelligible than talkers without such disorders (Borrie *et al.*, 2013). Studies have also shown differences in intelligibility across talker types, including familiar vs unfamiliar talkers (Levi, 2015; Nygaard and Pisoni, 1998) and familiar vs unfamiliar accents (Bent and Bradlow, 2003; Chung and Bong, 2021; Fuse *et al.*, 2018). These studies on talker and

accent familiarity have also demonstrated that relatively short exposure can improve intelligibility.

Intelligibility also captures differences between listener populations. For example, listeners with hearing impairment tend to perform worse than their normal hearing counterparts on intelligibility tasks in a variety of listening environments and with a variety of talkers (e.g., Suter, 1985). Further, intelligibility measures can capture some aspects of development that change between listener populations. For example, older adults tend to demonstrate reduced intelligibility as compared to younger adults, even when hearing status is controlled for (e.g., Dubno *et al.*, 2002; Rajan and Cainer, 2008). Similarly, younger children tend to perform more poorly than older children on tests of speech intelligibility (Elliott, 1979; Corbin *et al.*, 2016; Koopmans *et al.*, 2018).

Intelligibility measures also allow investigators to explore the effects of top-down information in speech perception and the relative contributions of top-down and bottom-up information. For example, higher predictability sentences (*the color of a lemon is yellow*) tend to be more intelligible than low predictability sentences (e.g., *mom thinks that it is yellow*; Kalikow *et al.*, 1977). This predictability benefit is modulated by age such that older adults rely more heavily on top-down semantic information (e.g., Pichora-Fuller *et al.*, 1995), whereas children rely more heavily on phonetic (or bottom-up) information (e.g., Elliott, 1979; Nittrouer and Boothroyd, 1990). Similarly, matching written context cues improve intelligibility compared to mismatching cues (Zekveld *et al.*, 2011). Further, items with high lexical frequency are more intelligible (more likely to be correctly reported) than low frequency items, across a range of listening populations, although this may reflect a response bias for high frequency items.³ Intelligibility is also impacted by a range of non-linguistic factors, including social information signaled by pictures (e.g., Babel and Russell, 2015; Hanulíková, 2021). Using intelligibility as an objective measure has helped us better understand how these top-down effects may emerge or shift as a function of other linguistic information, non-linguistic factors, listening environments, or listening populations (e.g., Baese-Berk *et al.*, 2021).

An additional benefit of intelligibility measures is that they capture individual variability. While above we have discussed population-level phenomena, individuals demonstrate variability on intelligibility tasks. This variability can be helpful when trying to investigate sources of differences in intelligibility or why particular results may emerge. For example, this amount of variability allows for investigations of whether various cognitive measures predict performance in different listening conditions (Bent *et al.*, 2016; Borrie *et al.*, 2017; Levi *et al.*, 2019; McLaughlin *et al.*, 2018).

Intelligibility scores are also subject to change over time. Listeners' scores tend to improve with practice at the task of transcription (e.g., Bradlow and Bent, 2008) and via exposure to a variety of types of speech [e.g., multiple unfamiliar accents (Baese-Berk *et al.*, 2013), dysarthric speech

(Borrie *et al.*, 2013), and sine-wave speech (Remez *et al.*, 1981)] among other types of listening challenge. This has allowed for investigations of what types of exposure are most effective at eliciting changes in intelligibility and have helped us better understand some of the cognitive processes underlying adaptation to unfamiliar speech.

Finally, on a practical level, intelligibility tests are easy to implement. It is possible to conduct the tests in person or via remote setups, and no specialized equipment is required. Further, the task is flexible for different populations, allowing either written or spoken responses. Scoring such responses is also easy: either the word is correctly transcribed, or it is not, allowing for a relatively simple analysis. A test, such as intelligibility, that can be done in a variety of settings, with a variety of populations, and requires no special equipment for administration or scoring is significantly easier compared to other techniques (described in Sec. IV).

Taken together, it is clear why intelligibility has been used for decades as a measure of speech perception, especially in investigations of challenging listening situations. However, there are also a number of challenges and limitations for using these measures, which we delineate below.

III. CHALLENGES AND LIMITATIONS OF INTELLIGIBILITY MEASURES

While initial assessment of intelligibility measures highlights the ease of scoring the data, a closer look reveals that this question is quite complicated. For example, many studies do not score transcriptions of all words in a response, but rather only the “keywords.” The definition of keywords is unclear, however. Some studies only use content words (nouns, verbs, adjectives, and adverbs); however, others also include (some) pronouns or prepositions and only exclude articles. Some studies allow for changes in transcription of tense for regularly conjugated verbs (*walk* for *walked* is scored as correct), while others require that tense be correctly transcribed for the word to be scored as correct. The same issue emerges for agreement of nouns [see Hustad (2006) and Miller (2013) for discussion].

An additional challenge is how to assess spelling differences from the target. For example, if a participant responds with a homophonous answer, it is unclear whether that should be marked as correct or incorrect, especially if it might impact interpretation of whether the participant actually understood the sentence. For example, “latter” for “ladder” may be an acceptable spelling, or it may be indicative that a participant has not understood the sentence in an example like “*the boy climbed the tall ladder.*” A related issue is how to code for known dialect differences, such as whether to count “Don” (/dɒn/) for “Dawn” (/dɔːn/) or *vice versa*. The spelling issue is also problematic if participants are being asked to transcribe unfamiliar speech or speech from a less familiar language. How much deviation is acceptable to the researcher? These issues allow for a large degree of individual experimenter freedom. Indeed, one commonly used tool for automatically scoring intelligibility data allows for researchers to determine which

variables they would like to manipulate (e.g., tense rules, plural rules, only scoring root words, which spelling mistakes are allowable, etc.; Borrie *et al.*, 2019).

This leads to a further complication, which is whether all mistakes are equal. While most can agree that *cat* is more similar to a target *cats* than *camp*, under many experimental scoring standards, both would be scored as incorrect, as would a failure to respond at all. This has led many to wonder whether a more fine-grained tool that allows for comparison along some dimension of similarity of phonemes or even of phonetic features may be preferable to the categorical “right” or “wrong” answers that are typically used in this type of data (Case *et al.*, 2018a,b). Some methods of scoring attempt to handle this to some degree by using a “fuzzy string matching” tool (e.g., Bosker, 2021). However, most of these tools use orthography rather than phonology, which can penalize spelling errors that some researchers might find acceptable and does not necessarily allow for the types of similarity comparisons that are more appropriate for speech sounds as opposed to written words (e.g., *ladder* vs *latter*). Taken together, it is clear that many decisions about the actual scoring (whether only exact matches are counted as correct, or whether additional credit is given for phonological overlap or morphologically related items) must be made by researchers. Furthermore, probing errors—such as examining differences in patterns between phonological and morphological/semantic errors—provides additional information about processing that classic intelligibility measures miss.

As described above, researchers may differ in whether they score certain whole words as correct or incorrect (e.g., Don-Dawn or latter-ladder) or differ in the level of coding (e.g., whole word vs phoneme). Some studies have used multiple coding schemes (e.g., Case *et al.*, 2018a,b; Levi, 2015) and found a similar pattern of results, suggesting that such decisions may not significantly alter research outcomes. However, we do not know of studies that have systematically examined this question. It is also likely that some coding decisions would alter the results of a speech perception task differently for different listener populations (e.g., coding root morphemes in verbs in past tense for individuals with developmental language disorder vs children with typical language development).

It is also unclear how to interpret cases where listeners fail to transcribe or repeat anything. While this could mean that listeners were unable to understand the words that were said in order to write them down or repeat them, it could also signal attention lapses, equipment malfunction, or other technical issues. Differentiating between these many possibilities is quite difficult, as all errors are treated identically, and all correct responses are treated as the same regardless of whether they are the result of a guess or a certain response. Although this issue is not limited to tests of intelligibility, it is nonetheless important to point out that researchers must make a decision about how to deal with non-responses.

An additional methodological difference across studies of intelligibility is whether listeners are asked to repeat or

write/type their responses. Verbal responses have the benefit of being usable for people with less knowledge of the written system (e.g., younger children) or individuals who may have poorer spelling or typing skills, which eliminates the coding concern surrounding items such as *latter-ladder*. One drawback of verbal responses is that they typically entail additional labor to transcribe the responses. In addition, verbal responses require researchers to accurately recognize the listeners' intended response (e.g., decide whether they are saying *Don* or *Dawn* or *pin* or *pen*). Writing or typing also takes more time and cognitive effort than repeating items verbally. As with some of the other methodological considerations noted above related to scoring, we are unaware of whether any study has examined whether the modality of response (written vs verbal) would alter the pattern of results.

Depending on the research question, the issue of uncertainty in what drives a particular correct or incorrect response could also lead to the misinterpretation of a result. This is perhaps especially true in circumstances where listeners are less familiar with the particular language variety of a speaker. For example, if a speaker repairs consonant clusters by inserting a vowel, but the "new" word with the vowel renders the sentence ungrammatical, the listener may alter the sentence structure in their response to make it grammatical, even if it no longer matches what the talker produced. Related to the issue of grammaticality, research has shown that in a sentence recognition task with children listening in quiet, knowledge of the syntactic structure and content words, whose role is more syntactic than semantic, has a greater impact on sentence recognition than semantics (Polišenská *et al.*, 2015). Thus, listeners may be trying to fit their responses into a syntactic frame that results in a grammatical utterance, even if this is not what the target was.

Additionally, it is worth noting that most studies that examine intelligibility must make the listening task more difficult, typically by adding some type of noise to the signal, because, as was pointed out above, listeners are very good at the task of recovering the message, even when it differs quite significantly from the variety(ies) they are most familiar with. Indeed, some commonly cited effects only emerge in more challenging signal-to-noise ratios. For example, the effect of the language of background talkers in babble may only emerge when the signal-to-noise ratio is quite hard (Van Engen and Bradlow, 2007). Further, by placing speech in noise to assess intelligibility, we may be assuming a linear relationship between the inclusion of noise and intelligibility that does not exist [see, e.g., Naylor (2016)]. This leads to a question of what intelligibility measures are telling us if they must be administered in unfavorable signal-to-noise ratios to demonstrate differences between conditions. That is, it is almost impossible to disentangle which aspects of the results of a particular study are attributable to the challenges of speech in noise and which may be attributed to challenges resulting from properties of a particular speaker or listener. What these studies are actually measuring is the interaction between noise and the

object of interest (e.g., non-native speech or a specific listener population), but many do not explicitly acknowledge this, instead assuming that their results only speak to the object of interest for a study.

Similarly, listeners are also sometimes asked to complete transcription or repetition tasks that measure intelligibility even if they do not understand any of the speech or if the speech makes little sense. For example, listeners may transcribe Jabberwocky sentences (i.e., sentences with normal syntactic structure but with nouns, verbs, adjectives, and adverbs replaced with non-words). Similarly, studies that use low predictability sentences in which listeners may perceive the individual words, but whose *meaning* may be awkward (e.g., *he can't consider the crib*; Kalikow *et al.*, 1977) or may border on nonsensical (e.g., *drums pour tall pets*; Stelmachowicz *et al.*, 2000) also demonstrate a differentiation between intelligibility (reporting words correctly) and understanding. This further demonstrates the difference between correctly reporting individual words being heard and comprehension or meaning of what a listener reports hearing. Relatedly, top-down contributions to intelligibility are reduced, although not non-existent, when listeners do not have contextual information (e.g., single word transcription).

Even in cases where a listener is asked to transcribe only meaningful sentences consisting of real words, many participants will occasionally transcribe anomalous sentences or non-words. Given this, it is unclear whether all listeners interpret the task in the same way. Some listeners may understand the instructions to "write down what you hear" to mean "write down all and only the real words that could be part of a coherent sentence," whereas others may interpret it to mean "write down what you think the speaker is trying to convey" or "write down what you think the speaker has articulated, including any misarticulations or errors." These strategies would result in different results for an intelligibility test and could exist within the same subjects in the same populations. While this issue could be addressed with highly specific instructions, it is still possible that listeners could interpret instructions differently and could perform slightly different tasks from one another.⁴

Perhaps the largest challenge with intelligibility measures is that the measure is uninformative about *what* specifically has gone awry during processing. This is in large part because intelligibility measures are almost exclusively offline, providing little information about real-time processing. Most studies do not combine reaction time and intelligibility measures, instead focusing only on the offline measure of words correctly transcribed. Further, unless the experiment specifically manipulates these factors, the influence of acoustic information is conflated with lexical, syntactic, and semantic knowledge, as well as other top-down factors that may impact processing.

IV. COMPLEMENTARY TASKS TO SIMPLE MEASURES OF INTELLIGIBILITY

Intelligibility, as described above, has both benefits and limitations. One of the primary limitations is that it is

unclear what, precisely, is being measured by intelligibility and how this measure of end point behavior may (or may not) correspond to other aspects of processing. Here, we describe two facets of processing, listening effort and higher-level processing, that do not necessarily correspond to intelligibility measures and discuss how measuring these may, in some cases, be more informative than using intelligibility measures alone. Specifically, the measures below allow researchers to address two separable but related problems: ceiling effects in intelligibility tasks, providing additional information about *what* a listener has misperceived, and understanding *why* the listener has misperceived.

It has long been assumed that speech that is less intelligible is also more effortful to process in general [see Van Engen and Peelle (2014) for a review]. That is, when a listening situation is challenging, a listener must recruit more cognitive resources to understand speech (Peelle, 2018; Rönnberg *et al.*, 2008; Rönnberg *et al.*, 2013; Rönnberg *et al.*, 2021). Aspects of listening effort can be measured using a variety of behavioral and physiological measures. Interestingly, however, intelligibility and effort are not strictly correlated. At a very basic level, this is clear because intelligibility and other measures, including comprehensibility (e.g., a subjective measure of how challenging it is to understand speech), do not strictly correspond to one another (e.g., Munro and Derwing, 1995a). Other studies examine effort more systematically and have demonstrated that fully intelligible speech produced by non-native talkers requires more effort to understand than equally intelligible speech from native talkers (McLaughlin and Van Engen, 2020). In this study, investigators used pupil dilation as a physiological measure of effort. Participants' pupils dilated more when listening to an unfamiliar accent than a familiar one, even though they were able to transcribe all of the speech accurately. This difference in speech processing would not have been captured by an intelligibility measure alone. An experiment using a dual-task paradigm also demonstrated that unfamiliar accents require more effort than familiar accents, even when the speech is fully intelligible (Brown *et al.*, 2020). This suggests that both physiological and cognitive measures of effort are sensitive to listening challenges in ways that intelligibility alone may not be—at a minimum, they provide us with additional methods of examining processing not available in classic intelligibility tasks because of ceiling effects.

Critically, even in cases where intelligibility is not at ceiling, listening effort measures may provide additional information about how a listener is processing speech. For example, Winn and Teece (2021) examined different types of “correct” and “incorrect” responses in an intelligibility task using pupillometry as the measure of effort. Correct responses that required some sort of correction (e.g., a section of the sentence was masked by noise) resulted in increased effort compared to correct responses that did not require the listener to correct their response. Similarly, errors that resulted in a semantically coherent (but incorrect) sentence (e.g., *the bride wore a white gown* for the target *the*

bride wore a white veil) resulted in increased effort compared to correctly transcribed sentences.

The effort measures described above are all more sophisticated than an additional, very basic, measure of effortful processing—reaction time. For decades, it has been clear that even when individuals are very good at decoding the signal at a basic level (analogous with intelligibility), they take more time to complete tasks in these sorts of challenging listening situations. For example, in speeded classification, listeners are highly accurate at identifying the initial consonant of a word, but they are slower when exposed to a switch in talkers (Mullennix and Pisoni, 1990). Similarly, unfamiliar accents result in increased processing time compared to familiar accents (Munro and Derwing, 1995b). Indeed, any challenges to particular tasks can impact accuracy and reaction time differently. Even though listeners accurately perceive words with ambiguous initial phonemes, this ambiguity increases looking time and latency during perception (McMurray *et al.*, 2002; Pisoni and Tash, 1974).

As described above, effort can be indexed both physiologically (using, e.g., pupillometry, heart rate variability, or skin conductance) and behaviorally (e.g., by measuring response times or performance on concurrent but secondary tasks). However, it is also possible that effortful processing can be indexed using other tools as well. For example, increased effort may result in intelligibility measures that are at ceiling (i.e., “perfect” performance in intelligibility), but that effort can cascade to other levels of processing, including memory and comprehension, which are rarely assessed in measures of intelligibility [see Van Engen and Peelle (2014) for a review].

Previous studies examining measures of intelligibility (e.g., word recognition or sentence recognition) have demonstrated that practice or familiarity can improve *intelligibility* as measured by more words correct (e.g., Baese-Berk *et al.*, 2013; Bent and Bradlow, 2003; Levi *et al.*, 2011; Van Engen, 2012). These studies often suggest that this benefit of intelligibility in the speech perception domain—reporting the word(s) that is spoken—could cascade to benefits in other domains by freeing up cognitive resources that would have been used at the level of *speech perception*. This idea, however, that benefits in the perceptual domain impact other levels of process has been minimally explored. The studies mentioned above on listening effort suggest that even when speech is reported correctly, listeners exert different levels of effort. For successful real-world communication, listeners not only need to perceive individual words, but must also remember information across utterances, interpret the meaning of words and phrases, and tie this information with stored long-term semantic information.⁵ Comprehension, for example, can be tested at different levels, including information recall (recalling one piece of information), information integration (combining two pieces of information), and inference (using information to make a prediction or implication; e.g., Sommers *et al.*, 2011). This is a critically important skill because substantial communication occurs outside of what is “said” strictly speaking.

Two skills beyond speech perception that could shed light on the impact of intelligibility and effort are memory and comprehension. Storing content in verbal working memory is critically important for speech perception because speech perception and comprehension require a listener to integrate information over a variety of time scales. Acoustic degradation of a speech signal has been shown to reduce recall of word pairs (Heinrich and Schneider, 2011) and word lists (Rabbitt, 1968; Cousins *et al.*, 2014). Similarly, when listening to unfamiliar speech, a listener is faced with more ambiguity and a less-interpretable signal than in ideal conditions. This uncertainty taxes working memory, as more resources are needed to understand the speech signal itself, leaving fewer cognitive resources for higher-level processing (Cowan, 1988; Cowan and Alloway, 1997; Nusbaum and Magnuson, 1997; Nusbaum and Schwab, 1986). Challenging listening situations do result in reduced memory for speech, while improving clarity of speech facilitates memory (e.g., Van Engen *et al.*, 2012).

Similarly, comprehension can be reduced in challenging listening situations. Both subjective measures of comprehension and objective measures demonstrate that listeners have more trouble *understanding* unfamiliar speech, even in cases where they are accurately able to transcribe it (Anderson Hsieh and Koehler, 1988; Major *et al.*, 2002; Munro and Derwing, 1995b). This is a critically important aspect of speech perception that is not addressed by basic measures of intelligibility.

Taken together, these findings suggest that speech perception in challenging listening conditions is a more complex construct than what intelligibility alone can capture. On their own, intelligibility data fail to account for issues of downstream processing (e.g., memory and comprehension) or for causes of these listening challenges (e.g., listening effort). Because listener responses are coded as binary (right or wrong), we may misinterpret results as being driven by the same processes because intelligibility scores are similar, even in cases where the same behavioral result may be driven by different processes. This problem is particularly concerning because often intelligibility measures are used to compare across listener groups (e.g., Pichora-Fuller *et al.*, 1995; Suter, 1985) or types of unfamiliar speech (e.g., Bent *et al.*, 2016; Borrie *et al.*, 2017; McLaughlin *et al.*, 2018) and to investigate change over time on particular tasks (e.g., Baese-Berk *et al.*, 2013; Borrie *et al.*, 2013; Bradlow and Bent, 2008). While intelligibility provides important information about speech recognition, other metrics are required to paint a complete picture of this complex behavior. Below, we describe recommendations for future work in this area.

V. CONCLUSIONS AND FUTURE DIRECTIONS

As discussed above, intelligibility is a useful and appealing measure on many dimensions. It is easy to implement. It captures researchers' intuitions about which circumstances should be challenging. It allows for measurement across a variety of populations. Of course, most researchers

recognize that the explanatory power of intelligibility measures is limited. While we can determine if a listener correctly transcribed a word or failed to do so, we cannot determine *why* they succeeded or failed. Therefore, if the goal of our work is to understand speech perception in challenging listening situations, we must not limit ourselves to this single metric.

If researchers choose to use intelligibility measures, they may want to consider more sophisticated analyses than the "whole word correct" approach that is often used. It is possible that fuzzy string matching tools (Bosker, 2021) could provide some nuance in the data. However, in addition to providing increased nuance, different scoring tools may emphasize different aspects of perception. For example, Felker *et al.* (2019) compare a variety of scoring measures, demonstrating that the choice of scoring metric emphasizes different features of perception. Future work could also compare how various ways of "counting" errors may impact findings (i.e., do the results of a study change when various ways of counting errors are compared). Further, when reporting data in this area, researchers could catalog (or at least provide samples of) the types of errors made, as previous work suggests different types of errors may be driven by different factors and may have different cognitive consequences (Winn and Teece, 2021). It is important to note, however, that while more fine-grained measures may be more sophisticated on one hand, they may also increase challenges on other dimensions. As an anonymous reviewer noted, because units within words (e.g., phonemes) are not independent, this may increase the statistical challenge of detecting true difference scores; this reviewer suggests that perhaps the easiest way to solve the problem of dependency among levels of representation (e.g., sounds, words, and phrases) would be to score at the sentence level, as each sentence is, theoretically, independent of the next in cases where the target is a sentence. However, in scoring at the sentence level, researchers would lose even more nuance than in the typically used "words correct" measures. This point further highlights the challenges in scoring and analyzing intelligibility measures and underscores the need for critical assessment of our tools and analyses.

It is also helpful to couple offline measures with measures that provide insight into real-time processing, such as online and reaction time measures. However, it is also possible to investigate similar questions about speech perception in challenging circumstances with other measures, such as tracking responses with mouse-tracking and eye-tracking (e.g., Hanulíková and Weber, 2012).⁶ Some work suggests that ERP measurements could also be informative in this area (e.g., Hanulíková *et al.*, 2012). For example, speech intelligibility is strongly correlated with electrophysiological measures (e.g., Vanthornhout *et al.*, 2018). A combination of these sorts of online, real-time measures with offline processing measures (e.g., a typed or verbal response) may be informative, especially when investigating processing in real time and trying to determine what aspects of processing might be impacted by various factors. That is, a more nuanced picture

of how offline measures correspond to these online processing measures could provide additional insight into many questions researchers often ask with offline measures alone.

Further, by investigating measures of effort and how these relate to intelligibility tasks, we will have a better sense of not only real-time processing, but when and how listeners may face and overcome challenges they encounter in the speech stream. Similarly, measures of downstream processing, including memory and comprehension, could supplement current intelligibility work to improve our understanding of speech perception in challenging listening situations.

Finally, researchers could take steps to begin to differentiate acoustic influences on intelligibility from lexical, syntactic, or semantic influences. Many intelligibility studies do not include acoustic information about their stimuli. While researchers may state that an unfamiliar accent deviates from a familiar norm, the exact aspects of those deviations often remain underspecified. Therefore, the actual source of the challenge for listeners is unclear. Our understanding of speech perception would be strengthened by further describing acoustic properties of stimuli (e.g., a speaker's vowel space area, speaking rate, etc.) or engaging in open science practices such that other researchers can investigate the acoustics and how those correspond with perception measures that make some speech more intelligible than others. We note that there are studies that examine individual acoustic properties on speech perception (e.g., Smiljanić and Bradlow, 2009). Further, speakers who are more or less intelligible than each other may change with different types of noise (Bent et al., 2009). However, these studies tend to examine or manipulate a single property rather than examining the acoustics of a person's speech more holistically.

In conclusion, we have demonstrated in this position piece that the classic idea of intelligibility, despite its many benefits, is a measure of speech perception that deserves the reconsideration that it has recently received in the speech science literature. Intelligibility measures have been helpful for elucidating many issues across decades of speech perception research and certainly will continue to be helpful in the future. However, these measures alone fail to capture the complexities of speech perception, especially in challenging listening situations. Researchers investigating these issues, and readers of their papers, can consider the myriad new tools, both methodological and statistical, that will provide more insight into the processing challenges faced by listeners in many real-world settings.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation (NSF) Grant No. BCS-2020805 to S.V.L. and M.M.B.B., a James S. McDonnell Foundation Opportunity Award to M.M.B.B., and NSF Grant No. BCS-2146993 to K.J.V.

¹We have purposely chosen to refer to these conditions as *challenging* instead of *adverse* as some studies do [see, e.g., Mattys et al. (2012)]. From our perspective, *adverse* refers to difficulties that are related to the

signal or to the context. We believe that *challenging* is more agnostic to the source of difficulties, which we believe is important since talkers from minoritized backgrounds are often blamed for misunderstandings, even when such misunderstandings can be driven, in part at least, by listener attitudes or experience [see, e.g., Baese-Berk et al. (2020) for a discussion of this issue with regard to non-native speech in particular].

²While this may be seen as a benefit in some subfields (i.e., those most interested in learning or adaptation over time), this could also be seen as a drawback for other fields (e.g., hearing sciences) that value test-retest reliability, or lack of change over time.

³It should be noted that these frequency effects can actually reverse for listening populations who are unfamiliar with a specific dialect or accent, especially if they are less familiar with reduction patterns typical for high frequency words (Levy et al., 2019).

⁴An anonymous reviewer notes that this problem is not inherent to intelligibility tasks and is more an issue with overly general or vague instructions. While it is true this problem could occur in other types of tests, it does seem that intelligibility tasks are particularly susceptible to this challenge given that listeners may have a different tolerance for uncertainty. That is, even if instructions state, "Write down all and only the words you are certain you understand," some listeners may be more comfortable than others with guessing and may rate their own certainty as higher than listeners who do not have a high threshold for comfort in guessing.

⁵An anonymous reviewer notes that this is a highly simplified and selective description of real-world communication; we agree.

⁶It should be noted that many existing studies using eye-tracking and event-related potential (ERP) measures in similar ways to current intelligibility tests also use written words as stimuli, which may complicate the comparison to more classic intelligibility tasks.

- Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). "Comprehension of familiar and unfamiliar native accents under adverse listening conditions," *J. Exp. Psychol. Hum. Percept. Perform.* **35**(2), 520–529.
- Anderson Hsieh, J., and Koehler, K. (1988). "The effect of foreign accent and speaking rate on native speaker comprehension," *Lang. Learn.* **38**(4), 561–613.
- Arbogast, T. L., Mason, C. R., and Kidd, G., Jr. (2002). "The effect of spatial separation on informational and energetic masking of speech," *J. Acoust. Soc. Am.* **112**(5), 2086–2098.
- Babel, M., and Russell, J. (2015). "Expectations and speech intelligibility," *J. Acoust. Soc. Am.* **137**(5), 2823–2833.
- Baese-Berk, M. M., Bent, T., and Walker, K. (2021). "Semantic predictability and adaptation to nonnative speech," *JASA Express Lett.* **1**(1), 015207.
- Baese-Berk, M. M., Bradlow, A. R., and Wright, B. A. (2013). "Accent-independent adaptation to foreign accented speech," *J. Acoust. Soc. Am.* **133**(3), EL174–EL180.
- Baese-Berk, M. M., McLaughlin, D. J., and McGowan, K. B. (2020). "Perception of non-native speech," *Lang. Linguist. Compass* **14**(7), e12375.
- Bent, T., Baese-Berk, M., Borrie, S. A., and McKee, M. (2016). "Individual differences in the perception of regional, nonnative, and disordered speech varieties," *J. Acoust. Soc. Am.* **140**(5), 3775–3786.
- Bent, T., and Bradlow, A. R. (2003). "The interlanguage speech intelligibility benefit," *J. Acoust. Soc. Am.* **114**(3), 1600–1610.
- Bent, T., Buchwald, A., and Pisoni, D. B. (2009). "Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech," *J. Acoust. Soc. Am.* **126**(5), 2660–2669.
- Borrie, S. A., Baese-Berk, M., Van Engen, K., and Bent, T. (2017). "A relationship between processing speech in noise and dysarthric speech," *J. Acoust. Soc. Am.* **141**(6), 4660–4667.
- Borrie, S. A., Barrett, T. S., and Yoho, S. E. (2019). "Autoscore: An open-source automated tool for scoring listener perception of speech," *J. Acoust. Soc. Am.* **145**(1), 392–399.
- Borrie, S. A., McAuliffe, M. J., O'Beirne, G. A., and Anderson, T. J. (2013). "The role of linguistic and indexical information in improved recognition of dysarthric speech," *J. Acoust. Soc. Am.* **133**(1), 474–482.
- Bosker, H. R. (2021). "Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies," *Behav. Res.* **53**(5), 1945–1953.

- Bradlow, A. R., and Bent, T. (2008). "Perceptual adaptation to non-native speech," *Cognition* **106**(2), 707–729.
- Brouwer, S., Van Engen, K. J., Calandruccio, L., and Bradlow, A. R. (2012). "Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content," *J. Acoust. Soc. Am.* **131**(2), 1449–1464.
- Brown, V. A., McLaughlin, D. J., Strand, J. F., and Van Engen, K. J. (2020). "Rapid adaptation to fully intelligible nonnative-accented speech reduces listening effort," *Q. J. Exp. Psychol.* **73**(9), 1431–1443.
- Calandruccio, L., Dhar, S., and Bradlow, A. R. (2010). "Speech-on-speech masking with variable access to the linguistic content of the masker speech," *J. Acoust. Soc. Am.* **128**(2), 860–869.
- Case, J., Seyfarth, S., and Levi, S. V. (2018a). "Short-term implicit voice-learning leads to a Familiar Talker Advantage: The role of encoding specificity," *J. Acoust. Soc. Am.* **144**(6), EL497–EL502.
- Case, J., Seyfarth, S., and Levi, S. V. (2018b). "Does implicit voice learning improve spoken language processing? Implications for clinical practice," *J. Speech Lang. Hear. Res.* **61**(5), 1251–1260.
- Cherry, E. C. (1953). "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.* **25**(5), 975–979.
- Chung, B., and Bong, H. K. M. (2021). "Intelligibility of Korean-accented English: Effects of listener's familiarity," *English Teach.* **76**(1), 33–56.
- Cooke, M., and Garcia Lecumberri, M. L. (2018). "Effects of exposure to noise during perceptual training of non-native language sounds," *J. Acoust. Soc. Am.* **143**(5), 2602–2610.
- Corbin, N. E., Bonino, A. Y., Buss, E., and Leibold, L. J. (2016). "Development of open-set word recognition in children: Speech-shaped noise and two-talker speech maskers," *Ear Hear.* **37**(1), 55–63.
- Cousins, K. A. Q., Dar, H., Wingfield, A., and Miller, P. (2014). "Acoustic masking disrupts time-dependent mechanisms of memory encoding in word-list recall," *Mem. Cogn.* **42**, 622–638.
- Cowan, N. (1988). "Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system," *Psychol. Bull.* **104**(2), 163–191.
- Cowan, N., and Alloway, T. (1997). "The development of working memory," *The Development of Memory in Childhood* (Taylor & Francis, London), pp. 163–199.
- Crum, D. (1974). "The effects of noise, reverberation, and speaker-to-listener distance on speech understanding," Ph.D. dissertation, Northwestern University, Evanston, IL.
- Dubno, J. R., Horwitz, A. R., and Ahlstrom, J. B. (2002). "Benefit of modulated maskers for speech recognition by younger and older adults with normal hearing," *J. Acoust. Soc. Am.* **111**(6), 2897–2907.
- Elliott, L. L. (1979). "Performance of children aged 9 to 17 years on a test of speech intelligibility in noise using sentence material with controlled word predictability," *J. Acoust. Soc. Am.* **66**(3), 651–653.
- Felker, E. R., Ernestus, M. T. C., and Broersma, M. (2019). "Evaluating dictation task measures for the study of speech perception," in *Proceedings of the 19th International Congress of Phonetic Sciences*, August 5–9, Melbourne, Australia, pp. 383–387.
- Freyman, R. L., Balakrishnan, U., and Helfer, K. S. (2001). "Spatial release from informational masking in speech recognition," *J. Acoust. Soc. Am.* **109**(5), 2112–2122.
- Fuse, A., Navichkova, Y., and Alloggio, K. (2018). "Perception of intelligibility and qualities of non-native accented speakers," *J. Commun. Disord.* **71**, 37–51.
- Hanulíková, A. (2021). "Do faces speak volumes? Social expectations in speech comprehension and evaluation across three age groups," *PLoS One* **16**(10), e0259230.
- Hanulíková, A., Van Alphen, P. M., Van Goch, M. M., and Weber, A. (2012). "When one person's mistake is another's standard usage: The effect of foreign accent on syntactic processing," *J. Cogn. Neurosci.* **24**(4), 878–887.
- Hanulíková, A., and Weber, A. (2012). "Sink positive: Linguistic experience with th substitutions influences nonnative word recognition," *Atten. Percept. Psychophys.* **74**(3), 613–629.
- Heinrich, A., and Schneider, B. A. (2011). "Elucidating the effects of ageing on remembering perceptually distorted word pairs," *Q. J. Exp. Psychol.* **64**(1), 186–205.
- Helfer, K. S., and Freyman, R. L. (2014). "Stimulus and listener factors affecting age-related changes in competing speech perception," *J. Acoust. Soc. Am.* **136**(2), 748–759.
- Hustad, K. C. (2006). "A closer look at transcription intelligibility for speakers with dysarthria: Evaluation of scoring paradigms and linguistic errors made by listeners," *Am. J. Speech Lang. Pathol.* **15**(3), 268–277.
- Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (1977). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," *J. Acoust. Soc. Am.* **61**(5), 1337–1351.
- Koopmans, W. J., Goverts, S. T., and Smits, C. (2018). "Speech recognition abilities in normal-hearing children 4 to 12 years of age in stationary and interrupted noise," *Ear Hear.* **39**(6), 1091–1103.
- Lane, H. (1963). "Foreign accent and speech distortion," *J. Acoust. Soc. Am.* **35**(4), 451–453.
- Levi, S. V. (2015). "Talker familiarity and spoken word recognition in school-age children," *J. Child Lang.* **42**(4), 843–872.
- Levi, S. V., Harel, D., and Schwartz, R. G. (2019). "Language ability and the familiar talker advantage: Generalizing to unfamiliar talkers is what matters," *J. Speech Lang. Hear. Res.* **62**(5), 1427–1436.
- Levi, S. V., Winters, S. J., and Pisoni, D. B. (2011). "Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible?," *J. Acoust. Soc. Am.* **130**(6), 4053–4062.
- Levy, H., Konieczny, L., and Hanulíková, A. (2019). "Processing of unfamiliar accents in monolingual and bilingual children: Effects of type and amount of accent experience," *J. Child Lang.* **46**(2), 368–392.
- Major, R. C., Fitzmaurice, S. F., Bunta, F., and Balasubramanian, C. (2002). "The effects of nonnative accents on listening comprehension: Implications for ESL assessment," *TESOL Q.* **36**(2), 173–190.
- Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). "Speech recognition in adverse conditions: A review," *Lang. Cogn. Process.* **27**(7–8), 953–978.
- Mayo, L. H., Florentine, M., and Buus, S. (1997). "Age of second-language acquisition and perception of speech in noise," *J. Speech. Lang. Hear. Res.* **40**(3), 686–693.
- McDermott, J. H. (2009). "The cocktail party problem," *Curr. Biol.* **19**(22), R1024–R1027.
- McLaughlin, D. J., Baese-Berk, M. M., Bent, T., Borrie, S. A., and Van Engen, K. J. (2018). "Coping with adversity: Individual differences in the perception of noisy and accented speech," *Atten. Percept. Psychophys.* **80**(6), 1559–1570.
- McLaughlin, D. J., and Van Engen, K. J. (2020). "Task-evoked pupil response for accurately recognized accented speech," *J. Acoust. Soc. Am.* **147**(2), EL151–EL156.
- McMurray, B., Tanenhaus, M. K., and Aslin, R. N. (2002). "Gradient effects of within-category phonetic variation on lexical access," *Cognition* **86**(2), B33–B42.
- Miller, N. (2013). "Measuring up to speech intelligibility," *Int. J. Lang. Commun. Disord.* **48**(6), 601–612.
- Mullennix, J., and Pisoni, D. B. (1990). "Stimulus variability and processing dependencies in speech perception," *Percept. Psychophys.* **47**(4), 379–390.
- Munro, M. J., and Derwing, T. M. (1995a). "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Lang. Learn.* **45**(1), 73–97.
- Munro, M. J., and Derwing, T. M. (1995b). "Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech," *Lang. Speech* **38**(3), 289–306.
- Nabelek, A. K., and Pickett, J. M. (1974). "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.* **17**(4), 724–739.
- Naylor, G. (2016). "Theoretical issues of validity in the measurement of aided speech reception threshold in noise for comparing nonlinear hearing aid systems," *J. Am. Acad. Audiol.* **27**(07), 504–514.
- Nittrouer, S., and Boothroyd, A. (1990). "Context effects in phoneme and word recognition by young children and older adults," *J. Acoust. Soc. Am.* **87**(6), 2705–2715.
- Nusbaum, H. C., and Magnuson, J. S. (1997). "Talker normalization: Phonetic constancy as a cognitive process," in *Talker Variability and Speech Processing* (Academic, New York), pp. 109–132.
- Nusbaum, H. C., and Schwab, E. C. (1986). "The role of attention and active processing in speech perception," in *Pattern Recognition by Humans and Machines* (Elsevier, New York), pp. 113–157.
- Nygaard, L. C., and Pisoni, D. B. (1998). "Talker-specific learning in speech perception," *Percept. Psychophys.* **60**(3), 355–376.

- Peelle, J. E. (2018). "Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior," *Ear Hear.* **39**(2), 204–214.
- Pichora-Fuller, M. K., Schneider, B. A., and Daneman, M. (1995). "How young and old adults listen to and remember speech in noise," *J. Acoust. Soc. Am.* **97**(1), 593–608.
- Pisoni, D. B., and Tash, J. (1974). "Reaction times to comparisons within and across phonetic categories," *Percept. Psychophys.* **15**(2), 285–290.
- Polišenská, K., Chiat, S., and Roy, P. (2015). "Sentence repetition: What does the task measure?," *Int. J. Lang. Commun. Disord.* **50**(1), 106–118.
- Rabbitt, P. M. (1968). "Channel-capacity, intelligibility and immediate memory," *Q. J. Exp. Psychol.* **20**(3), 241–248.
- Rajan, R., and Cainer, K. E. (2008). "Aging without hearing loss or cognitive impairment causes a decrease in speech intelligibility only in informational maskers," *Neuroscience* **154**(2), 784–795.
- Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). "Speech perception without traditional speech cues," *Science* **212**(4497), 947–949.
- Rönnberg, J., Holmer, E., and Rudner, M. (2021). "Cognitive hearing science: Three memory systems, two approaches, and the ease of language understanding model," *J. Speech. Lang. Hear. Res.* **64**(2), 359–370.
- Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., and Rudner, M. (2013). "The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances," *Front. Syst. Neurosci.* **7**, 31.
- Rönnberg, J., Rudner, M., Foo, C., and Lunner, T. (2008). "Cognition counts: A working memory system for ease of language understanding (ELU)," *Int. J. Audiol.* **47**(sup2), S99–S105.
- Smiljanić, R., and Bradlow, A. R. (2009). "Speaking and hearing clearly: Talker and listener factors in speaking style changes," *Lang. Linguist. Compass.* **3**(1), 236–264.
- Sommers, M. S., Hale, S., Myerson, J., Rose, N., Tye-Murray, N., and Spehar, B. (2011). "Listening comprehension across the adult lifespan," *Ear Hear.* **32**(6), 775–781.
- Stelmachowicz, P. G., Hoover, B. M., Lewis, D. E., Kortekaas, R. W. L., and Pittman, A. L. (2000). "The relation between stimulus context, speech audibility, and perception for normal-hearing and hearing-impaired children," *J. Speech. Lang. Hear. Res.* **43**(4), 902–914.
- Suter, A. H. (1985). "Speech recognition in noise by individuals with mild hearing impairments," *J. Acoust. Soc. Am.* **78**(3), 887–900.
- Van Engen, K. J. (2010). "Similarity and familiarity: Second language sentence recognition in first- and second-language multi-talker babble," *Speech Commun.* **52**(11–12), 943–953.
- Van Engen, K. J., and Bradlow, A. R. (2007). "Sentence recognition in native- and foreign-language multi-talker background noise," *J. Acoust. Soc. Am.* **121**(1), 519–526.
- Van Engen, K. J., Chandrasekaran, B., and Smiljanic, R. (2012). "Effects of speech clarity on recognition memory for spoken sentences," *PLoS One* **7**(9), e43753.
- Van Engen, K. J., and Peelle, J. E. (2014). "Listening effort and accented speech," *Front. Hum. Neurosci.* **8**, 577.
- Van Engen, K. J., Phelps, J. E., Smiljanic, R., and Chandrasekaran, B. (2014). "Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker," *J. Speech. Lang. Hear. Res.* **57**(5), 1908–1918.
- Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., and Francart, T. (2018). "Speech intelligibility predicted from neural entrainment of the speech envelope," *J. Assoc. Res. Otolaryngol.* **19**(2), 181–191.
- Van Wijngaarden, S. J., Steeneken, H. J., and Houtgast, T. (2002). "Quantifying the intelligibility of speech in noise for non-native listeners," *J. Acoust. Soc. Am.* **111**(4), 1906–1916.
- Vermiglio, A. J., Herring, C. C., Heeke, P., Post, C. E., and Fang, X. (2019). "Sentence recognition in steady-state speech-shaped noise versus four-talker babble," *J. Am. Acad. Audiol.* **30**(1), 54–65.
- Winn, M. B., and Teece, K. H. (2021). "Listening effort is not the same as speech intelligibility score," *Trends Hear.* **25**, 23312165211027688.
- Zekveld, A. A., Rudner, M., Johnsrude, I. S., Festen, J. M., Van Beek, J. H., and Rönnberg, J. (2011). "The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise," *Ear Hear.* **32**(6), e16–e25.