ASHA American Speech-Language-Hearing Association

JSLHR

Research Note

Spread the Word: Enhancing Replicability of Speech Research Through Stimulus Sharing

Julia F. Strand[a] and **Violet A. Brown**[b]

[a] Department of Psychology, Carleton College, Northfield, MN [b] Department of Psychological & Brain Sciences, Washington University in St. Louis, MO

ARTICLE INFO

Article History:
Received May 12, 2022
Revision received August 3, 2022
Accepted October 26, 2022

Editor-in-Chief: Cara E. Stepp
Editor: Rachel M. Theodore

https://doi.org/10.1044/2022_JSLHR-22-00267

ABSTRACT

Purpose: The ongoing replication crisis within and beyond psychology has revealed the numerous ways in which flexibility in the research process can affect study outcomes. In speech research, examples of these "researcher degrees of freedom" include the particular syllables, words, or sentences presented; the talkers who produce the stimuli and the instructions given to them; the population tested; whether and how stimuli are matched on amplitude; the type of masking noise used and its presentation level; and many others. In this research note, we argue that even seemingly minor methodological choices have the potential to affect study outcomes. To that end, we present a reanalysis of six existing data sets on spoken word identification in noise to assess how differences in talkers, stimulus processing, masking type, and listeners affect identification accuracy.
Conclusions: Our reanalysis revealed relatively low correlations among word identification rates across studies. The data suggest that some of the seemingly innocuous methodological details that differ across studies—details that cannot possibly be reported in text given the idiosyncrasies inherent to speech—introduce unknown variability that may affect replicability of our findings. We therefore argue that publicly sharing stimuli is a crucial step toward improved replicability in speech research.
Supplemental Material: https://doi.org/10.23641/asha.21985907

Researchers make many methodological choices in the course of an experiment. These "researcher degrees of freedom" (Simmons et al., 2011; Wicherts et al., 2016) include which tasks to use, what population to draw from, how many trials to present, which observations are considered outliers, and many others. Simmons et al. (2011) suggest that the flexibility inherent to the research process—coupled with publication bias (Rosenthal, 1979)—contributes to the presence of false positives in the literature. These false positives may partially account for the recent replication crisis in psychology (see Open Science Collaboration, 2015) and other sciences (e.g., Ioannidis, 2005; Loscalzo, 2012).

Although researcher degrees of freedom are present in all realms of research, the number of independent methodological decisions required when creating and presenting speech stimuli makes this issue particularly pertinent to speech researchers. These decisions may or may not affect study outcomes, but they certainly introduce variability of unknown magnitude and potential bias. For example, stimuli in speech perception experiments (isolated words, sentences, etc.) necessarily represent only a small subset of all available stimuli in a language, often leading to somewhat arbitrary decisions regarding the particular words or sentences to be included in a study.[1] However, item selection is just one of many decisions speech perception researchers must make when designing a study; we must also decide which talker(s) will produce the stimuli, what type of masking noise to use, the signal-to-noise ratio

Correspondence to Julia F. Strand: jstrand@carleton.edu. **Publisher Note:** This article is part of the Forum: Promoting Reproducibility for the Speech, Language, and Hearing Sciences. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

[1] Note that the arbitrary nature of stimulus selection is certainly not true of all experiments in our field. Indeed, the authors of this research note have conducted experiments in which the pool of possible stimuli was quite limited due to experimental constraints (see Strand et al., 2017).

(SNR) at which to present stimuli, the speaking rate, speaking clarity, and so on (decisions that must be made even when those variables are not directly relevant to the research question). In addition to the choices regarding stimulus selection and audio recording, researchers must also decide whether to remove ambient background noise and/or equate the amplitude of auditory stimuli and must select specific algorithms for doing so. Thus, the particular methodology that a researcher implements is just one of many combinations of methodological decisions that might have been selected.

Researchers are typically careful to use a consistent set of experimental conditions *within* an experiment—indeed, from a methodological perspective, it is desirable to reduce variability that can be attributed to sources other than the effect of interest. However, seemingly innocuous methodological decisions such as those described above may introduce unexpected (and unaccounted for) sources of variability *across* experiments that may affect the replicability of results, particularly if the experimental effect is small or highly context-dependent. As a result, if two researchers attempt to study the same phenomenon but their methods differ, it is not clear whether any conflicting results were due to a false positive or negative, or instead to methodological differences and context-dependent effects. That is, the conflicting findings may both be "true" and may reflect the bounds of the effect.

It is well established that some methodological choices can substantially influence experimental outcomes. For example, performance decrements as a result of decreases in SNR are more pronounced for steady-state noise than for two-talker babble (Brungart, 2001); thus, if an effect depends on differences in SNR, the choice of background noise type is likely to affect outcomes. As another example, one of the most robust findings in the speech perception literature is that seeing a talker, in addition to hearing their voice, benefits intelligibility, but this effect is much more pronounced at difficult SNRs (Sumby & Pollack, 1954); thus, the magnitude of any audiovisual benefit effects will depend on the particular noise level the researchers decide to use. However, other seemingly minor researcher degrees of freedom may also affect outcomes in unforeseen ways and therefore impede replicability.

In this research note, we argue two things are necessary to increase the robustness and replicability of results in speech research. First, researchers should include more comprehensive information about how stimuli are processed and presented (see the Additional Recommendations section). However, even the most detailed reporting about stimuli cannot capture information about idiosyncratic features of the talker (and doing so would require pages of text that detract from the main point of the experiment). Therefore, our second more critical recommendation is that researchers share the audio or video files they used in their research.

It is relatively common practice to include a list of the stimuli that were used in a given study (e.g., in an Appendix section), and databases exist for sharing speech stimuli for reuse (Bradlow, n.d.; Brown, 2020). However, sharing the exact stimulus files that were used in a particular study is relatively rare in our discipline. For example, in 2021, only seven of the papers published in *Journal of Speech, Language, and Hearing Research (JSLHR)* made reference to stimulus sharing: five by linking to a repository that contained the stimuli and two by noting that the stimuli were available "upon request from the authors."[2]

To shed light on whether and how methodological choices regarding stimulus creation affect word identification accuracy across data sets, we analyzed data from six prior studies on spoken word recognition in noise. If minor methodological choices—such as the talker who produces the speech[3] or the method for processing the audio recordings—are innocuous, then identification accuracy for the same words should be highly correlated across data sets (i.e., influenced only by measurement error and therefore bound by the reliability of the measures). Thus, attenuations in these correlations (beyond the attenuation attributable to less-than-perfect measurement reliability) reflect the extent to which methodological choices affect study outcomes.

Reanalysis of Existing Data

The six data sets we analyzed included three from previous work by one of the authors (J.F.S.) and three that were obtained from collaborators or by contacting the authors of the studies.[4] In all six studies, normal-hearing participants attempted to identify spoken words in noise, presented either in isolation or in a consistent carrier phrase such as "the word is ...." Raw data and analysis code, as well as stimuli for studies conducted by J.F.S., are available at https://osf.io/v38ej/. Table 1 provides a description of the six data sets (some grouped by listener group or SNR).

First, we calculated the average accuracy with which each target word was correctly identified in each data set. We then calculated correlations among the (scaled, log-odds) identification rates for the words in the 11 (sub)data

[2]We opted not to present this as a percentage of the total papers given that some papers published in *JSLHR* do not have stimuli to share (e.g., meta-analyses and systematic reviews).

[3]Note that some features of talker selection (such as which dialect they speak) can be expected to substantially affect study outcomes; here, we refer to the choice of particular talker within the dialect being studied.

[4]These data sets were chosen because we were able to obtain item-level data from them and they contained enough overlapping items across data sets to run correlations among them.

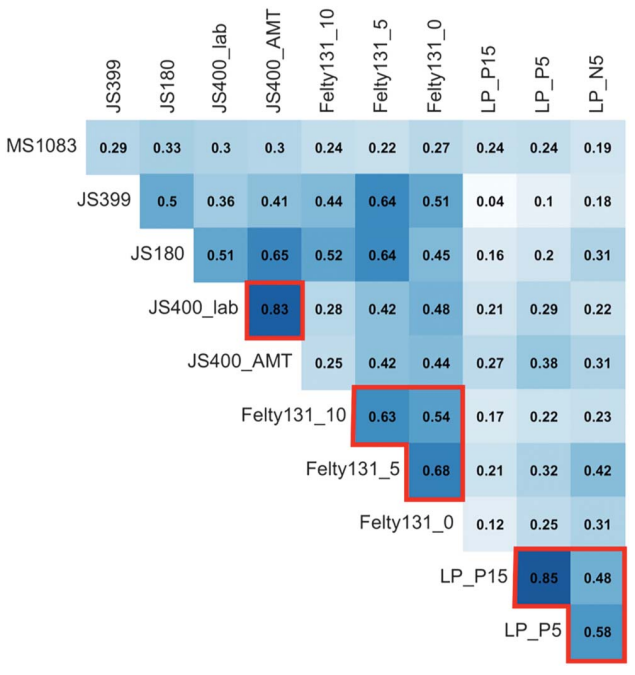
Table 1. Descriptions of six data sets.

Data set	Setting	SNR	Masker	*N* (words)	*N* (participants)	Average accuracy	Label
Sommers, unpublished	Lab	–5	6-talker babble	1,083	93 (heard ⅓ of words)	.44	MS1083
Strand, unpublished	Lab	–2	6-talker babble	399	50	.50	JS399
Strand & Sommers (2011)	Lab	–4	6-talker babble	180	72	.28	JS180
Slote & Strand (2016)	Lab[a]	0	6-talker babble	400	53	.54	JS400_lab
	Online	0	6-talker babble	400	96	.37	JS400_AMT
Felty et al. (2013)	Lab	10	6-talker babble	127	192 (heard ⅓ of words)	.72	Felty131_10
		5	6-talker babble	127	192 (heard ⅓ of words)	.53	Felty131_5
		0	6-talker babble	127	192 (heard ⅓ of words)	.29	Felty131_ 0
Luce & Pisoni (1998)	Lab	15	White noise	876	10	.65	LP_P15
		5	White noise	876	10	.47	LP_P5
		–5	White noise	876	10	.12	LP_N5

[a]This data set contains some missing observations; 30 of the 400 words in the JS400_lab data set were only presented to 17 participants, so not all average values in the data frame are divisible by 53.

sets (see Figure 1). Each value in Figure 1 represents the Pearson correlation coefficient between a pair of data sets, and correlations are based on words that were present in both data sets in the pair being compared (see Supplemental Material S1 for a corresponding scatter plot matrix). The median overlap between data sets was 119 words, the minimum was 17 words (comparing the JS400 data sets with the Felty data sets), and the maximum was 813 words (comparing the LP data sets with the MS1083 data set).

Figure 1. Pearson correlation coefficients among data sets. Darker colors represent stronger correlations, and those that are not significant ($p > .05$) have a white background. Red boxes indicate that the same recordings were used within the same study. Note that pairs of data sets differ in the number of words they have in common, which can lead to nonsignificant correlations with larger *r* values than significant ones.



Correlations among the six unique data sets (i.e., excluding those within the red boxes in Figure 1) ranged from $r = .04$ to $r = .65$. The median correlation among the different data sets was $r = .30$. However, it should be noted that some of the strongest of these correlations (the correlations of the three SNRs of the Felty data set with both JS399 and JS180, which ranged from $r = .44$ to $r = .64$) have the smallest number of overlapping items ($N = 17$ and $N = 33$ shared words), so the magnitude of those correlations should be interpreted cautiously. The Felty data sets had markedly fewer overlapping items with the two data sets listed above, as well as the JS400 data sets (17, 33, and 38 overlapping items, whereas the next lowest had 67 overlapping items). When we removed pairs with fewer than 38 overlapping items, the median correlation dropped to $r = .26$.

These results suggest that the extent to which particular words are easily identifiable is not consistent across studies, presumably as a result of characteristics of the recordings themselves, the speakers used, the nature of the masking noise, or the participants (see Toscano & Allen, 2014, for similar findings in phone recognition). The relatively low correlations between data sets are particularly concerning in the context of the ongoing replication crisis in psychology: If words are not recognized at relatively consistent rates from one experiment to the next, then small or context-dependent effects may not replicate across studies. In an effort to understand why the correlations among data sets were lower than might be expected, we identified four broad sources of systematic variability that may affect replicability in spoken word identification studies: characteristics of the talker, differences in digital stimulus processing, features of the masking noise, and characteristics of the listeners.

Variability Source #1: Characteristics of the Talker

Identification rates for particular words depend on the talker's style of speaking. In addition to group differences that are known to affect speech intelligibility—such as speaker sex (Yoho et al., 2019), differences in language background (Bent & Bradlow, 2003; Clarke & Garrett, 2004), and dialect variations (Clopper, 2021)—idiosyncrasies in speaking style that remain within the bounds of the accent and dialect with which the listener is most familiar can also affect speech identification (see McCloy et al., 2015). Indeed, speech produced in a clear speaking style is identified more accurately than speech produced conversationally (Van Engen, 2017; Van Engen et al., 2012), and speakers differ in the clarity of their speech (Smiljanić & Bradlow, 2005). Speakers also differ substantially in their rate of speech (Bond & Moore, 1994)—a feature that is known to affect speech intelligibility (Bradlow & Pisoni, 1999)—as well as the fundamental frequency and other vocal features (Bradlow et al., 1996) that may interact with the background noise and influence identification rates.

To assess how different talkers affect speech intelligibility across data sets, we examined correlations among word identification rates for each of the six talkers in the JS180 data set (Strand & Sommers, 2011). These correlations ranged from $r = .31$ to $r = .51$ (see Figure 2). The main difference among these data sets was the talkers used, so the correlations here suggest that a substantial amount of variation in word identification accuracy may be attributable to talker idiosyncrasies.[5] These small-to-moderate correlations are particularly striking given that people tend to speak more clearly than they otherwise would when producing recordings for speech perception experiments, which would be expected to reduce variability across talkers.

The relatively low correlations among identification rates for different talkers may lead to another issue related to replicability: If identification rates for the *same words* are not consistent across talkers, then lexical characteristics that are known to predict word recognition accuracy *across words* may also have varying effects across talkers.

[5]In JS180, every participant identified words produced by every talker but the specific words they heard from a given talker varied such that participants heard one sixth of the words produced by each talker. For example, one sixth of the participants heard "cat" produced by Talker 1 and one sixth of the participants heard "cat" produced by Talker 2 (see Strand & Sommers, 2011), meaning that the accuracies for each token were derived from different participants. However, participants from a single sample were randomly assigned to word-by-talker lists, therefore minimizing the likelihood that participant differences could be responsible for the small-to-moderate correlations across talkers.

Figure 2. Correlations among word identification accuracy of the six talkers in the JS180 data set.

To assess this possibility, we examined the relationship between word identification accuracy for each of the six talkers from JS180 and two commonly used lexical variables—frequency of occurrence (Brysbaert & New, 2009) and neighborhood density (Luce & Pisoni, 1998). Given that word frequency and neighborhood density tend to be positively correlated but have opposite effects on word identification accuracy (i.e., high-density words tend to occur more frequently in the language, but density impairs identification whereas frequency facilitates it), we assessed the effects of density and frequency on identification by calculating their semipartial correlations with identification accuracy. Semipartial correlations were evaluated using the *spcor* function in the "ppcor" package (Kim, 2015), and accuracy was defined as the proportion correct for a word produced by a given talker (collapsed across participants). Frequency values were obtained from Brysbaert and New (2009) and represent log-scaled word frequency values from a corpus of film and television subtitles. Density values were obtained from Balota et al. (2007) and represent the number of words that can be created by a single-phoneme substitution of the target. The semipartial correlation between frequency and identification accuracy, controlling for density, was significant for all talkers in the JS180 data set (see Table 2). In contrast, the effects of density, controlling for frequency, were more variable and only generated statistically significant semipartial correlations for two of the six talkers.

These results indicate that weak effects may or may not emerge depending on the talker who produces the

Table 2. Semipartial correlations between word identification accuracy and both frequency of occurrence (controlling for neighborhood density) and neighborhood density (controlling for word frequency).

Talker	Frequency	Density
Talker 1	$sr = .29$, $p < .001$	$sr = -.15$, $p = .047$
Talker 2	$sr = .28$, $p < .001$	$sr = -.10$, $p = .19$
Talker 3	$sr = .20$, $p = .007$	$sr = -.11$, $p = .14$
Talker 4	$sr = .26$, $p < .001$	$sr = -.16$, $p = .03$
Talker 5	$sr = .30$, $p < .001$	$sr = -.06$, $p = .40$
Talker 6	$sr = .27$, $p < .001$	$sr = .003$, $p = .97$

stimuli. Furthermore, the variability shown here is likely to represent the lower bound of what would be predicted across studies, given that in addition to using different talkers, studies are likely to differ in how they process the stimuli, the population sampled from, equipment used, and so on. Taken together, this suggests that the choice of which talker to use may significantly influence study outcomes. One potential solution to this issue is, when appropriate, to use multiple talkers in a study to ensure that any effects obtained are not conditional on the choice of talker.[6]

Variability Source #2: Digital Stimulus Processing

When researchers generate auditory stimuli for research on spoken word identification, they must make several decisions about whether and how to process the stimuli. For example, researchers may choose to perform some form of leveling to ensure that words do not differ substantially in the amplitude at which they are presented. Most researchers do not report their method for leveling auditory stimuli, and those who do typically equate stimuli on root-mean-square (RMS) amplitude using either Adobe Audition (Tye-Murray et al., 2010, 2016, 2015) or Praat (Van Engen et al., 2014, 2017). However, Audition has numerous options for leveling audio (e.g., ITU-R BS.1770-3, which uses K-weighting/perceived loudness, total RMS, peak amplitude), and after selecting one of them, the user has control over various parameters—including the target loudness, tolerance, and maximum true peak level for ITU-R BS.1770-3, and target loudness for total RMS[7]; researchers rarely, if ever, report these details.

Researchers must also decide whether or not they should remove ambient noise (e.g., low-level hum from electrical appliances) from the audio files. Although the standard noise reduction process is similar in various audio processing softwares (i.e., a noise sample is taken from a portion of the audio file during which speech is not present, then the amplitude of those frequencies is selectively reduced throughout the audio track), researchers can adjust the parameters to vary the extent to which ambient noise is removed. For example, both Adobe Audition and Audacity allow the user to specify the amplitude reduction of the background noise, the number of frequency smoothing bands, and the algorithm's sensitivity to noise, but Audition allows for precise control over several other parameters that are not available in Audacity (including fast Fourier transform size, spectral decay rate, etc.). Depending on the magnitude and nature of the effect of interest, these decisions regarding signal processing may or may not affect outcomes. However, the additional variability introduced by differences in stimulus processing may limit the replicability of fragile effects, and sharing materials (and specifying details regarding stimulus processing; see below) enables researchers to identify the particular sources of variability that may account for these differences across studies.

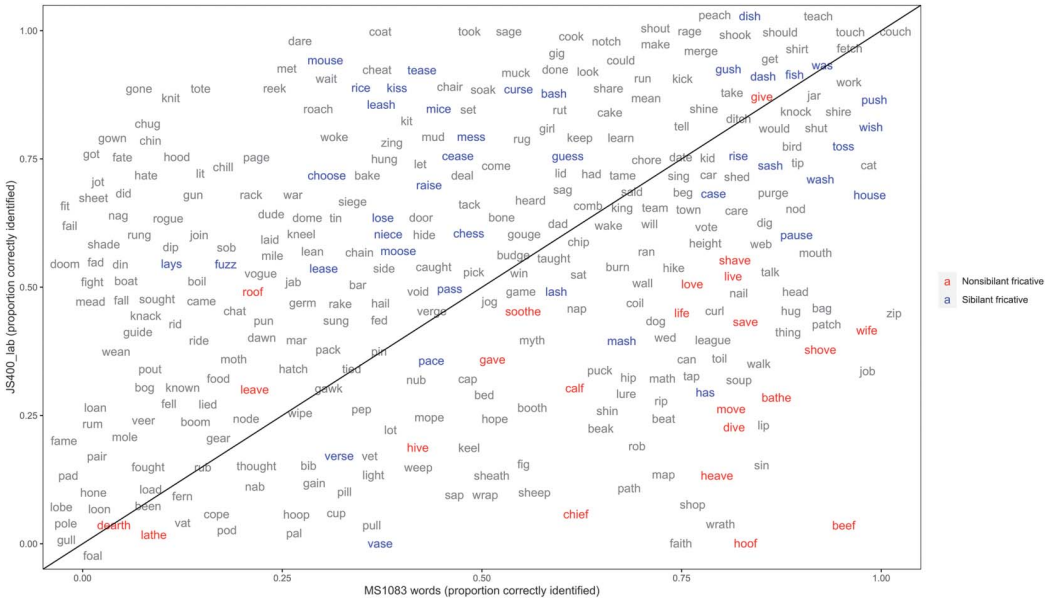
The available data do not allow us to explicitly test the effects of different signal processing decisions on word recognition accuracy because the processing techniques in these studies are unknown. However, a scatter plot of identification rates in the two largest data sets that used the same background noise and SNR demonstrates some consistent patterns in identification rates; indeed, visual inspection of scatter plots revealed that some types of words in one data set were systematically under- or overestimated in the other (see Figure 3). For example, only 9% of the words ending in nonsibilant fricatives (e.g., "hoof" and "faith"; shown in red) were identified at higher rates in the JS400 data set than in the MS1083 data set, whereas 61% of the words ending in sibilant fricatives (e.g., "was" and "push"; shown in blue) were identified at higher rates in JS400 (see Figure 3).

These discrepancies may be the result of speaker idiosyncrasies (e.g., perhaps the talker in JS400 produces nonsibilants in a less intelligible manner or more rapidly than the talker in MS1083) or differences in the particulars of the masking noise—although both studies used six-talker babble, the babble files were constructed using different talkers and different sentences. Alternatively or in addition, the differences may be a function of the signal-processing techniques the researchers employed. For example, sibilants (e.g., /s/) tend to have higher overall amplitude than nonsibilants (e.g., /θ/); thus, a study that matches entire words on a target RMS level will tend to increase the amplitude of the rest of the word more when it contains a nonsibilant than when it contains a sibilant. For example, if the /s/ in "sick" has higher amplitude than the /θ/ in "thick" but the /ik/ portions are the same

[6]When making decisions about using one or multiple talkers, researchers should consider how perceptual learning of talker identity may affect outcomes (see Nygaard & Pisoni, 1998).

[7]Although these may seem like minor details, they can be influential: For example, too much amplification can lead to clipping, which distorts the speech and may impair intelligibility.

Figure 3. Scatter plot showing word identification accuracy in the JS400 and MS1083 data sets. The black line represents $y = x$. Labeling of sibilant and nonsibilant fricatives corresponds to the coda consonant.



amplitude, matching those words on RMS level would tend to increase the amplitude of the /ik/ in "thick" more than in "sick." If another study does not match words on target RMS amplitude but instead opts to present the words at the amplitude at which they were recorded, the ends of those words would be equivalent in amplitude. Thus, discrepancies in signal processing may account for differences in identification rates for the same words across studies.

Variability Source #3: Masking Noise

Differences in spoken word identification rates across studies may also be attributable to the characteristics of the masking noise used. For example, the fact that Luce and Pisoni (1998) presented speech in white noise whereas all other studies used six-talker babble may account for the somewhat lower correlations between the Luce and Pisoni data sets and the others. However, the effects of masker type cannot be directly assessed with the data we present here because none of the studies held all other factors constant (e.g., participant group, SNR, and testing setting) while manipulating masker type.

In addition to differences in the type of masker used (e.g., white noise vs. six-talker babble), differences in the acoustic features of maskers of the same type are also likely to produce different levels of interference. If, for example, one study used six-talker babble produced by talkers whose voices were very similar to the target (e.g., same gender and age) and another used babble that was more dissimilar to the target, these different six-talker babble files are likely to produce different levels of interference. Indeed, although both the MS1083 and JS400 data sets used six-talker babble, the relatively low correlation in identification rates between the two studies may be partially attributable to the fact that the babble was produced by different talkers and may have been leveled or mixed in different ways.

These findings suggest that the characteristics of the maskers—in addition to traits of the talkers and the stimuli—may affect study outcomes. Regardless of the type of noise employed, we encourage researchers to share the maskers used in their studies in addition to the target speech files. Sharing masking noise enables other researchers to directly assess the influence of various types of masking noise on identification rates for a particular set of words.

Variability Source #4: Listener Differences

Another source of variability that may contribute to low correlations across studies is differences in the listeners included in the sample. Indeed, no two studies include exactly the same participants, even if the samples come from similar participant pools. This sampling variability—particularly when paired with small sample sizes—may produce different identification rates for the same words across studies. Only one pair of experiments in the current investigation can speak to this possibility: The JS400_lab and JS400_AMT data sets (Slote &

Strand, 2016) used the same words, speakers, and recordings, but the JS400_lab data set was collected in a typical laboratory setting, whereas the JS400_AMT data set was collected online using Amazon's Mechanical Turk. Identification rates across these data sets were highly correlated (r = .83)—despite different participants and variability in listening environments—and approached the split-half reliabilities of the two experiments (r = .91 for in-lab, r = .94 for online). This suggests that identification rates using the same recordings are relatively stable across participants and provides reassurance for issues of replicability given that listeners will always be different between experiments and labs. The fact that correlations are highest when the same recordings of target words are used suggests that features of the particular audio recordings may be more important than previously thought.

A core feature of scientific research is that results can be independently replicated. The data presented here demonstrate that what researchers may have assumed were innocuous methodological choices may actually affect study outcomes, which as a consequence may threaten replicability in our discipline. Reporting features of the stimuli and experimental design that are known or expected to affect study outcomes is commonplace in the literature. However, given the infinite number of idiosyncrasies inherent in speech that cannot possibly be reported, we recommend sharing study materials like speech stimuli (along with data and code) in an accessible repository as a helpful step toward increasing transparency in research (see Klein et al., 2018).

Benefits of Sharing Stimuli

Sharing stimuli has multiple benefits. First, it facilitates future attempts to replicate and/or extend previous work. For example, if one paper describes a finding within a sample of young adults, future researchers can easily assess whether the finding applies to older adults as well if they have access to the original stimuli. Thus, stimulus sharing enables researchers to explicitly test whether characteristics of the listeners (or the nature of the testing environment, headphones used, etc.) affect study outcomes while holding other features of the experiment constant.

Sharing stimuli can also help "protect" the researchers doing the sharing from subsequent failures to replicate. For example, if a finding does not replicate, the replication team can rerun the experiment using the original stimuli and compare the results to those obtained using their own stimuli, but only if the original stimuli are publicly available. This may shed light on whether a failure to replicate is driven by differences in stimulus materials (e.g., using different talkers) or some other factor (e.g., differences in the sample, statistical power, etc.).

The existence of stimulus repositories has pedagogical benefits as well. Indeed, replication studies represent an efficient and accessible entry point into research for undergraduates given that the theoretical background, methods, and analyses that are relevant to the experiment are already clearly defined. Thus, publicly available stimuli make the task of incorporating replication attempts into undergraduate curricula less daunting for faculty who either run undergraduate research labs—in which students may join the lab with little or no research experience, and turnover rates are often high so it is important that students learn about research methods as quickly as possible—or include a lab component in their undergraduate courses (e.g., see Wagge et al., 2019).

Some researchers may be reluctant to share stimuli given the cost of audio equipment and time spent recording and editing; indeed, stimulus creation represents a significant investment of time and resources. However, sharing materials—and stimuli in particular—can help reduce barriers for researchers with fewer resources and thereby address issues related to equity in our discipline. In addition, having access to high-quality stimuli enables early career researchers who do not yet have established labs to begin data collection more quickly than they otherwise might be able to. Broadly speaking, sharing materials increases accessibility in science and builds inclusivity in our discipline (Ledgerwood et al., 2022).

How to Share Stimuli

There are many options for sharing stimuli in a publicly accessible way. The most straightforward is to do so in an existing repository such as the Open Science Framework (http://osf.io), which provides a stable link and DOI that can be embedded directly in the paper. Using an existing repository is preferable to linking to stimuli on a researcher's personal website because personally maintained websites are more likely to change (e.g., if a researcher moves institutions) and are therefore less stable over time. The least accessible method of sharing is to make materials available "by request." This shifts the burden onto future researchers and increases the likelihood that the stimuli will not be accessible because the original researchers changed their e-mail addresses, forgot where they stored the files, or were simply too busy to respond (see Savage & Vickers, 2009, for evidence of low rates of compliance for data sharing).

Regardless of where stimuli are shared, it is important that the stimuli are not only available, but also decipherable—that is, presented in a format that makes them easy for others to understand and use. Although there exist many guidelines for sharing data and code (e.g., Wilkinson et al., 2016), there are fewer guidelines for sharing materials, possibly because subdisciplinary

differences make it difficult to generate standardized guidelines. In the case of speech materials, at minimum it must be clear to a naïve reader what the stimulus files represent (e.g., which word/sentence/experimental condition a particular stimulus file corresponds to) and how they are organized. We therefore recommend creating a document that is posted along with the stimuli that contains a full list of the stimuli, notes about file-naming conventions (e.g., "cat_A.wav" refers to the word "cat" produced by talker A), and any other details necessary for a naïve reader to comprehensively understand what each file contains. Furthermore, stimuli should be shared in commonly used formats (e.g., .wav files) so future researchers can easily open and manipulate them.

Although most researchers record stimuli in lossless (uncompressed) formats, it is sometimes necessary to present lossy (compressed) recordings in the experiment itself. For example, some online stimulus presentation software (e.g., Gorilla Experiment Builder; Anwyl-Irvine et al., 2020) requires that auditory files be uploaded in .mp3 (or .ogg) rather than .wav format. In these cases, we recommend that researchers post both the original lossless files and those that were actually used in the study and clearly label which is which.

It is advisable that researchers posting their stimuli publicly specify a license that gives others guidance about how the stimuli may be reused. There are several Creative Commons licenses (https://creativecommons.org/about/cclicenses/) that indicate whether the creator of the stimuli must be credited, whether the stimuli may be used for commercial purposes, and whether the stimuli may be edited/altered. We recommend the CC BY-NC license, which allows reusers to "distribute, remix, adapt, and build upon the material in any medium or format for noncommercial purposes only, and only so long as attribution is given to the creator." This ensures that the creator will be credited but allows subsequent researchers to edit the materials as needed. It is also important to describe the sharing plan to the talker producing the stimuli and obtain their consent at the time of recording. This can be achieved by asking the talker to sign a document affirming that they understand how and why the stimuli will be used and shared (akin to a participant signing a consent form).

We recommend that researchers share their stimuli at the time a manuscript is submitted to a journal or preprint server because this is when the details regarding stimulus creation, file structure, naming conventions, and so on are fresh in mind. However, researchers who have stimuli available to share from previous papers can still post those stimuli online for future use. One of the authors of this research note (V.A.B.) has created an Open Science Framework page (https://osf.io/nbwaj/) to facilitate sharing stimuli from published or unpublished work.

Additional Recommendations

In addition to sharing stimuli, we urge researchers to include even more detail about the nature of the stimuli and how they were recorded and processed in the Method section. Although it is common to describe the gender and language background of the talker (and background babble when relevant), it is rare to see descriptions of whether any noise removal was implemented, the software and settings used to level the amplitude of the stimuli, the format in which the recordings were presented (e.g., lossy or lossless), the speaking rate, the average duration of the words, and any other potentially relevant details. Note that several of these details (e.g., file duration) may not be necessary to include in the paper if stimuli are shared but are crucial if they are not.

We also encourage researchers to think about and explicitly comment on whether and how they believe their choice of stimulus materials is likely to affect outcomes in Constraints on Generality statements (Simons et al., 2017). These statements give authors the opportunity to state the situations in which they expect their findings to generalize (or not), which not only helps ensure that findings will not be overblown but may also protect authors from potential failures to replicate. For example, if the authors explicitly state that their findings may not apply to individuals with hearing loss and a later study fails to find the effect in that population, that lack of an effect is not seen as a failure to replicate, but rather as information about the bounds of the effect. Constraints on Generality statements often specify whether the effect is specific to a particular target population (e.g., Is a finding for normal-hearing participants likely to generalize to those with hearing loss?), type of stimulus materials (e.g., Is a finding on isolated words likely to extend to words embedded in context?), linguistic trait of the talker or listener (e.g., Is the finding robust to changes in dialect?), listening environment (e.g., Might changing the masking noise type or SNR affect the outcomes?), and so forth.

Rather than appealing to researchers to share their materials, some journals—including *Nature* (Springer Nature, n.d.) and *Psychological Science* (Lindsay, 2017)—have opted to make sharing data and materials a precondition of publication. Adopting this and other practices laid out in the Transparency and Openness Promotion guidelines (Nosek et al., 2015) can help journals ensure that the work they are publishing is as transparent as possible. *JSLHR* has made key steps toward encouraging greater transparency, including implementing Registered Reports and offering Open Science Badges. In addition to these important steps, explicitly encouraging researchers to share their data and materials will help facilitate the replicability of research in our discipline.

Conclusions

Sharing stimuli is not currently standard practice in speech research, and undisclosed variability in how speech stimuli are recorded and processed may lead to inconsistencies in the literature. We therefore encourage researchers to more thoroughly document their process regarding stimulus creation and publicly share those stimuli for others to reuse. It is important to note that we are not advocating that researchers use a more limited set of stimuli; indeed, assessing whether findings are robust to differences in talkers, stimulus processing choices, SNRs, and so on is an important component of our work. Rather, we are emphasizing that sharing stimuli has benefits in terms of facilitating replication, reconciling inconsistent results, and reducing barriers to access for researchers who have fewer resources. Thus, along with other practices such as preregistration (Nosek et al., 2018), publishing Registered Reports (Chambers, 2013), and sharing data and code (Houtkoop et al., 2018), stimulus sharing will help establish a more robust and replicable literature.

Data Availability Statement

Data are available at https://osf.io/v38ej/.

Acknowledgments

This work was supported by the National Science Foundation through a Graduate Research Fellowship awarded to Violet A. Brown (DGE-1745038) and the National Institute on Deafness and Communication Disorders via a grant to Julia F. Strand (R15-DC018114). The authors are grateful to Mitchell Sommers, Tom Gruenenfelder, and Paul Luce for sharing data sets.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods, 39*(3), 445–459. https://doi.org/10.3758/BF03193014

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America, 114*(3), 1600–1610. https://doi.org/10.1121/1.1603234

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic–phonetic characteristics of inadvertently clear speech. *Speech Communication, 14*(4), 325–337. https://doi.org/10.1016/0167-6393(94)90026-4

Bradlow, A. R. (n.d.). *SpeechBox*. https://speechbox.linguistics.northwestern.edu/#!/home

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America, 106*(4), 2074–2085. https://doi.org/10.1121/1.427952

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic–phonetic talker characteristics. *Speech Communication, 20*(3–4), 255–272. https://doi.org/10.1016/S0167-6393(96)00063-5

Brown, V. A. (2020). *Stimuli and other resources for speech researchers.* Open Science Framework. https://doi.org/10.17605/OSF.IO/NBWAJ

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America, 109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Clopper, C. G. (2021). Perception of dialect variation. In *The handbook of speech perception* (pp. 333–364). Wiley. https://doi.org/10.1002/9781119184096.ch13

Felty, R. A., Buchwald, A., Gruenenfelder, T. M., & Pisoni, D. B. (2013). Misperceptions of spoken words: Data from a random sample of American English words. *The Journal of the Acoustical Society of America, 134*(1), 572–585. https://doi.org/10.1121/1.4809540

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science, 1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Kim, S. (2015). ppcor: An R package for a fast calculation to semi-partial correlation coefficients. *Communications for Statistical Applications and Methods, 22*(6), 665–674. https://doi.org/10.5351/CSAM.2015.22.6.665

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Hofelich Mohr, A., Ijzerman, H., Nilsonne, G., Vanpaemel, W., & Frank, M. C. (2018). A practical guide for transparency in psychological science. *Collabra: Psychology, 4*(1), 20. https://doi.org/10.1525/collabra.158

Ledgerwood, A., Hudson, S.-K. T. J., Lewis, N. A., Jr., Maddox, K. B., Pickett, C. L., Remedios, J. D., Cheryan, S., Diekman, A. B., Dutra, N. B., Goh, J. X., Goodwin, S. A., Munakata, Y., Navarro, D. J., Onyeador, I. N., Srivastava, S., & Wilkins, C. L. (2022). The pandemic as a portal: Reimagining psychological science as truly open and inclusive. *Perspectives on Psychological Science, 17*(4), 17456916211036654. https://doi.org/10.1177/17456916211036654

Lindsay, D. S. (2017). Sharing data and materials in psychological science. *Psychological Science, 28*(6), 699–702. https://doi.org/10.1177/0956797617704015

Loscalzo, J. (2012). Irreproducible experimental results: Causes, (mis)interpretations, and consequences. *Circulation, 125*(10), 1211–1214. https://doi.org/10.1161/CIRCULATIONAHA.112.098244

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

McCloy, D. R., Wright, R. A., & Souza, P. E. (2015). Talker versus dialect effects on speech intelligibility: A symmetrical study. *Language and Speech, 58*(3), 371–386. https://doi.org/10.1177/0023830914559234

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., . . . Yarkoni, T. (2015). Scientific standards. Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America, 115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355–376. https://doi.org/10.3758/BF03206860

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi.org/10.1037//0033-2909.86.3.638

Savage, C. J., & Vickers, A. J. (2009). Empirical study of data sharing by authors publishing in PLoS journals. *PLOS ONE, 4*(9), Article e7078. https://doi.org/10.1371/journal.pone.0007078

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science, 12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Slote, J., & Strand, J. F. (2016). Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods, 48*(2), 553–566. https://doi.org/10.3758/s13428-015-0599-7

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America, 118*(3, Pt. 1), 1677–1688. https://doi.org/10.1121/1.2000788

Springer Nature. (n.d.). *Reporting standards and availability of data, materials, code and protocols.* Retrieved May 5, 2022, from https://www.nature.com/nature-portfolio/editorial-policies/reporting-standards

Strand, J. F., Brown, V. A., Brown, H. E., & Berg, J. J. (2017). Keep listening: Grammatical context reduces but does not eliminate activation of unexpected words. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 44,* 962–973. https://doi.org/10.1037/xlm0000488

Strand, J. F., & Sommers, M. S. (2011). Sizing up the competition: Quantifying the influence of the mental lexicon on auditory and visual spoken word recognition. *The Journal of the Acoustical Society of America, 130*(3), 1663–1672. https://doi.org/10.1121/1.3613930

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212–215. https://doi.org/10.1121/1.1907309

Toscano, J. C., & Allen, J. B. (2014). Across- and within-consonant errors for isolated syllables in noise. *Journal of Speech, Language, and Hearing Research, 57*(6), 2293–2307. https://doi.org/10.1044/2014_JSLHR-H-13-0244

Tye-Murray, N., Sommers, M. S., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing, 31*(5), 636–644. https://doi.org/10.1097/AUD.0b013e3181ddf7ff

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., & Sommers, M. S. (2015). The self-advantage in visual speech processing enhances audiovisual speech recognition in noise. *Psychonomic Bulletin & Review, 22*(4), 1048–1053. https://doi.org/10.3758/s13423-014-0774-3

Tye-Murray, N., Spehar, B. P., Myerson, J., Hale, S., & Sommers, M. S. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging, 31*(4), 380–389. https://doi.org/10.1037/pag0000094

Van Engen, K. J. (2017). Clear speech and lexical competition in younger and older adult listeners. *The Journal of the Acoustical Society of America, 142*(2), 1067–1077. https://doi.org/10.1121/1.4998708

Van Engen, K. J., Chandrasekaran, B., & Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PLOS ONE, 7*(9), Article e43753. https://doi.org/10.1371/journal.pone.0043753

Van Engen, K. J., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2014). Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language, and Hearing Research, 57*(5), 1908–1918. https://doi.org/10.1044/JSLHR-H-13-0076

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception & Psychophysics, 79*(2), 396–403. https://doi.org/10.3758/s13414-016-1238-9

Wagge, J. R., Brandt, M. J., Lazarevic, L. B., Legate, N., Christopherson, C., Wiggins, B., & Grahe, J. E. (2019). Publishing research with undergraduate students via replication work: The collaborative replications and education project. *Frontiers in Psychology, 10,* 247. https://doi.org/10.3389/fpsyg.2019.00247

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7,* 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., . . . Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*(1), 160018. https://doi.org/10.1038/sdata.2016.18

Yoho, S. E., Borrie, S. A., Barrett, T. S., & Whittaker, D. B. (2019). Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology. *Attention, Perception & Psychophysics, 81*(2), 558–570. https://doi.org/10.3758/s13414-018-1635-3