AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001640

Assessing the Effects of "Native Speaker" Status on Classic Findings in Speech Research

Julia F. Strand[1], Violet A. Brown[2], Katrina Sewell[1], Yuxin Lin[1], Emmett Lefkowitz[1], and Caroline G. Saksena[1]

[1] Department of Psychology, Carleton College

[2] Department of Psychological and Brain Sciences, Washington University in St. Louis

It is common practice in speech research to only sample participants who self-report being "native English speakers." Although there is research on differences in language processing between native and nonnative listeners (see Lecumberri et al., 2010, for a review), the majority of speech research that aims to establish general findings (e.g., testing models of spoken word recognition) only includes native speakers in their sample. Not only is the "native English speaker" criterion poorly defined, but it also excludes historically underrepresented groups from speech perception research, often without attention to whether this exclusion is likely to affect study outcomes. The purpose of this study is to empirically test whether and how using different inclusion criteria ("native English speakers" vs. "nonnative English speakers") affects several well-known phenomena in speech perception research. Five hundred participants completed word ($N = 200$) and sentence ($N = 300$) identification tasks in quiet and in moderate levels of background noise. Results indicate that multiple classic findings in speech perception research—including the effects of noise level, lexical density, and semantic context on speech intelligibility—persist regardless of "native English" speaking status. However, the magnitude of some of these effects differed across participant groups. Taken together, these results suggest that researchers should carefully consider whether native speaker status is likely to affect outcomes and make decisions about inclusion criteria on a study-by-study basis.

Public Significance Statement

Most research in experimental psychology is conducted on undergraduates enrolled in colleges and universities in the United States and other Western countries. Research on spoken language often further homogenizes samples to include only "native" speakers of the language. Here, we show that for some general research questions (i.e., those that do not specifically aim to address differences in language processing between individuals with different first languages), this restriction may not substantially influence study outcomes. In conducting this research, we hope to encourage researchers to carefully consider whether restricting samples to "native" speakers is necessary, thereby promoting inclusivity in research.

Keywords: spoken word recognition, speech perception, native speaker

Supplemental materials: https://doi.org/10.1037/xge0001640.supp

Joseph Toscano served as action editor.

Julia F. Strand https://orcid.org/0000-0001-5950-0139

All data, materials, and code are publicly available in the Open Science Framework and can be accessed at https://osf.io/tvwur/. Portions of this project were presented at the Auditory Perception, Cognition, and Action Meeting in 2022 as well as the Annual Meeting of the Psychonomic Society in 2023. The work was supported by Carleton College and the National Institute on Deafness and Other Communication Disorders, National Institutes of Health (Grant R15-DC018114), awarded to Julia F. Strand. The authors thank Naseem Dillman-Hasso, Eva Hadjiyanis, Chiamaka Ifedi, Zhaobin Li, Ellen Mamantov, Binny Onabolu, and Gigi Paulig for their helpful conversations.

Julia F. Strand played a lead role in conceptualization, funding acquisition, and supervision and an equal role in data curation, formal analysis, investigation, methodology, resources, visualization, writing–original draft, and writing–review and editing. Violet A. Brown played a lead role in formal analysis, a supporting role in project administration, and an equal role in conceptualization, investigation, methodology, visualization, writing–original draft, and writing–review and editing. Katrina Sewell played a lead role in investigation, a supporting role in conceptualization, data curation, formal analysis, visualization, and writing–review and editing, and an equal role in project administration. Yuxin Lin played a supporting role in conceptualization, formal analysis, investigation, methodology, writing–original draft, and writing–review and editing. Emmett Lefkowitz played a supporting role in conceptualization, methodology, and writing–original draft. Caroline G. Saksena played a supporting role in conceptualization, methodology, and writing–original draft.

Correspondence concerning this article should be addressed to Julia F. Strand, Department of Psychology, Carleton College, One North College Street, Northfield, MN 55057, United States, or Violet A. Brown, Department of Psychological and Brain Sciences, Washington University in St. Louis, 1 Brookings Drive Street, St. Louis, MO 63130, United States. Email: jstrand@carleton.edu or violetbrownpsych@gmail.com

One of the goals of psychological research is gaining a better understanding of human behavior, yet many studies use samples that do not represent the human population generally. Indeed, the majority of psychological research has been conducted on what is often referred to as "WEIRD" (Western, Educated, Industrialized, Rich, and Democratic) samples (Henrich et al., 2010; but see Syed, 2021, for arguments against using the WEIRD acronym, given that it does not incorporate important sources of diversity like race, ethnicity, religion, and class) and on undergraduate students in universities (Henrich et al., 2010; Rad et al., 2018). Although using nonrepresentative samples is quite common, many articles do not address whether and how the composition of their samples may affect their findings (Rad et al., 2018).

Limited sampling is common in many areas of psychology, but this issue is exacerbated in research on spoken language, as the sample is often further constrained to include only "native English speakers."[1] Indeed, of the 100 most cited behavioral studies on spoken word recognition and speech perception (excluding studies that had the explicit goal of testing effects of native speaker status or bilingualism), 62 of them indicated that their sample included only native speakers, and this choice was rarely justified (see Supplemental Materials for more information on the sampling process). Of the remaining 38 articles, three stated that participants were English speakers, and 35 made no mention of the language background of the participants. Thus, research with the goal of establishing general behavioral findings about how humans understand spoken language frequently restricts samples to native speakers, typically without explaining why.[2]

This is not to say that nonnative speakers are not represented in language research. Indeed, there is a rich literature on how native speaker status affects a host of phenomena related to processing speech (see Lecumberri et al., 2010). For example, previous research has shown that nonnative listeners tend to identify speech less accurately than native listeners, tend to be more negatively affected by background noise (Rogers et al., 2006) and lexical difficulty (Bradlow & Pisoni, 1999), and benefit less from seeing the talker (Xie et al., 2014) and from semantic contextual cues (Shi, 2014).[3] These differences may be driven by experience with the language; that is, lexical representations for nonnative listeners may be less tuned to English than those for native listeners (i.e., representations for listeners may not match the input as precisely; Flege, 1992), and this "mistuning" may persist through higher levels of linguistic processing (Bradlow & Alexander, 2007). This research on native versus nonnative listeners deliberately explores the effects of language background (including first language learned, proficiency, regularity of use, etc.) on speech-related outcomes, which necessitates careful attention to differences across participants. In contrast, research that aims to test general phenomena in speech perception (like the 100 articles mentioned above) typically only samples native listeners. This inclusion criterion is rarely justified, suggesting that the practice is driven by the conventions of the field rather than the anticipated effects of language background on study outcomes.

Although sampling only native speakers may afford experimental control, the practice poses some challenges. First, it treats linguistic variables (e.g., "monolingual" vs. "bilingual," "native" vs. "nonnative") as binary categories, despite the fact many people's language experiences defy such categorizations (see Luk, 2023; Weissler et al., 2023). Within this binary, the term "native speaker" lacks an agreed-upon definition (B. Brown et al., 2023; Cheng et al., 2021). Some studies base their definitions on an individual's history with a language (Abrahamsson & Hyltenstam, 2009), others center their definitions around proficiency (Stern, 1983), and in some cases the concept of nativeness depends on whether the participant's identity is tied to the language (Debenport, 2011). As a result, native speaker status is also operationalized differently across studies: Some studies measure language proficiency with the use of a language test such as the Test of English as a Foreign Language exam (Bradlow & Bent, 2002) or the Cambridge Advanced Examination (Cooke et al., 2008), others use various self-reported proficiency scales (Broersma & Scharenborg, 2010), and others simply ask the participant to affirm native speaker status without clear criteria about what that means. The lack of a widely agreed-upon definition means that a participant who may be deemed a native speaker in one study may be considered a nonnative speaker in another (Cheng et al., 2021), thereby leading to inconsistent findings across studies that aim to test the same effects. Furthermore, many articles do not report how "native" is defined in the study; of the 62 articles in our survey that restricted their samples based on language background, none explained what criteria they used to define native speaker status. This may make it difficult for future researchers to conduct direct replication attempts and could lead to "failed" replications simply because of sample differences.

In addition to the vagueness of the criterion, requiring that participants be native speakers reinforces hegemonic conceptions of how people use and learn language; it implies that the model organism for studying language had the specific acquisition experience of learning languages strictly sequentially (see Kutlu & Hayes-Harb, 2023, for a recent special issue on how normative practices in language research may perpetuate the marginalization of some groups). Given that research participants in the United States already tend to be limited to undergraduates in universities (Henrich et al., 2010; Rad et al., 2018), imposing additional constraints based on language background has the consequence of further homogenizing already limited samples. The decision to limit samples to native English speakers also means that the samples overrepresent White participants (U.S. Census Bureau, 2020), a trend that perpetuates existing inequities in the discipline (Ledgerwood et al., 2022). Indeed, in many areas of psychology, White samples are passively considered the neutral default, and race is primarily discussed only for non-White samples (Roberts & Mortenson, 2023).

Finally, for many general phenomena, the main effects of interest may emerge regardless of whether the sample is limited to native English speakers, and including more diverse samples increases the

[1] We acknowledge the limitations associated with using the term "native English speaker" (Cheng et al., 2021). In our literature review, we use that term when discussing how the concept has been used previously in the literature. In our approach, we opt instead to use the terminology L1 (for people who learned English before any other language) and LX (for people who learned another language first; see Dewaele, 2018 for arguments about the advantages of using this terminology).

[2] Although the practice of limiting samples to native speakers is prevalent in speech research, it also occurs in other domains, including research on visual search (e.g., Poole & Kane, 2009), syntactic priming (Bernolet et al., 2016), autobiographical memory (Berntsen et al., 2019), effects of visual context on memory (Gilmore et al., 2016), and the Stroop effect (Hatukai & Algom, 2017), to name a few.

[3] Note, however, that publication bias may inflate published estimates of differences between native and nonnative listeners (see de Bruin et al., 2015, for an example of publication bias in the bilingual cognitive advantage literature).

generalizability of the research. Many findings in speech perception research are present and robust in both listener groups: Both groups identify speech more accurately when they can see the talker's face (Xie et al., 2014), benefit from semantic context (Shi, 2014), show poorer speech intelligibility in more difficult levels of background noise (Bradlow & Bent, 2002; Cutler et al., 2004; Rogers et al., 2006), benefit from slower speech in difficult listening conditions (Hazan & Simpson, 2000), and show greater release from masking when the target speech and masker are produced by talkers of different genders (Cooke et al., 2008). Although the magnitude of some findings may be moderated by the language background of the listener (e.g., nonnative listeners tend to be more affected by background noise), these effects are typically much smaller than the main effects being studied (background noise), and those main effects operate similarly in both groups (noise impairs intelligibility for all listeners). Thus, although researchers may exclude nonnative listeners from their samples to facilitate experimental control, when it comes to establishing findings that could apply to all listeners (not just native English speakers), this practice hampers generalizability, reinforces historical inequities in research, and in some cases may only have a minor influence on study outcomes.

The Present Study

This study evaluates whether and how the composition of the sample affects well-known phenomena in speech perception research. Our goal is to shed light on an issue that is well-known in some subdisciplines of psychology (e.g., bilingualism research), but is often overlooked in others. This study tested speech identification accuracy for isolated words (Experiment 1) and semantically unconstrained and constrained sentences (Experiment 2). In each experiment, we include samples of L1 ("native") and LX ("nonnative") English speakers to enable us to (a) assess whether the pattern of results is consistent when the L1 and LX groups are analyzed separately and (b) determine whether the effects are more pronounced in one group than the other by statistically comparing the two groups. We defined L1 English speakers as participants who self-reported learning English before any other language, and LX English speakers as participants who self-reported learning any other language before English (Dewaele, 2018).

In both the word and sentence identification tasks, we presented speech in quiet and in a moderate level of background noise. Although increasing the level of the background noise impairs speech intelligibility for all listeners, this effect tends to be more pronounced for nonnative listeners (Cooke et al., 2008). Similarly, lexically difficult words—that is, those with high neighborhood density, low frequency of occurrence in the language, and/or high neighborhood frequency (Luce & Pisoni, 1998)—are identified less accurately than lexically easy words, but the effect of lexical difficulty tends to be more pronounced for nonnative relative to native listeners (Bradlow & Pisoni, 1999). Thus, we chose to manipulate background noise level and lexical difficulty because these variables robustly affect word identification accuracy across participant populations, but the effects differ in magnitude across L1 and LX listeners. Specifically, previous work has demonstrated that effects of noise and lexical difficulty tend to be *more* pronounced in LX listeners. To ensure that the variables included in this study elicit effects that interact with listener group in both directions, in Experiment 2 we manipulated a variable that typically produces *less* pronounced effects in LX listeners.

In the sentence identification task (Experiment 2), in addition to manipulating background noise level, we also varied the strength of the semantic cues provided by the early part of the sentence. Semantic context has been shown to improve speech intelligibility in native and nonnative listeners (Bradlow & Alexander, 2007); indeed, identification accuracy for the final word of a sentence is improved when the word is predictable based on the preceding context (e.g., "Let us decide by tossing a *coin*" vs. "Jane has a problem with the *coin*"). However, benefit from semantic contextual cues tends to be more pronounced in native relative to nonnative listeners (Bradlow & Alexander, 2007; Mayo et al., 1997; Shi, 2010, 2014), perhaps because nonnative listeners have less experience with the language and with the process of rapidly integrating top-down and bottom-up cues in their nonnative language. Previous work has demonstrated that differences in speech processing between native and nonnative listeners tend to be most pronounced when listeners are required to use both top-down and bottom-up cues (Bradlow & Alexander, 2007), so manipulating both semantic context and the level of the background noise enables us to assess differences in speech processing between the two groups in situations in which they are most likely to emerge.

Across Experiments 1 and 2, we manipulate variables that are expected to elicit both stronger (noise, lexical difficulty) and weaker (semantic constraint) effects in LX listeners. Varying the direction of potential interaction effects involving manipulated variables and listener group enables us to evaluate the robustness of these classic effects to variations in participant populations, even when varying the participant population might make it more difficult for the researcher to detect the effect. Additionally, by including one experiment using isolated words as stimuli (Experiment 1) and one experiment using semantically constrained and unconstrained sentences (Experiment 2), this study evaluates the extent to which the language background of the participants affects their accuracy at identifying speech in noise at multiple levels of linguistic processing. This is not to say, of course, that the findings we report below would apply to all experimental manipulations; indeed, there are certainly situations in which the language background of the listener might be expected to more substantially affect study outcomes (e.g., differences in voicing perception between English and Spanish speakers; see the General Discussion section for more on this). Rather, the manipulations included in this set of experiments enable us to conduct an initial investigation into the effects of language background on a subset of classic findings in speech perception research.

In addition to measuring word identification accuracy, we also assessed subjective listening effort to obtain a measure of the subjective experience associated with speech processing in L1 and LX listeners. Measuring listening effort—the cognitive resources necessary to comprehend speech (Pichora-Fuller et al., 2016)—provides valuable insight into a listener's experience that may be missed by measures of intelligibility alone. For example, previous work has shown that even when speech intelligibility is matched between L1 and LX listeners, LX listeners still expended greater listening effort to achieve the same level of performance (Borghini & Hazan, 2018). Therefore, collecting data on both intelligibility and effort provides a more holistic view of the challenges of a given listening task. Thus, the current work includes both replications of previous findings (e.g., the finding that LX listeners tend to be more affected by background noise than L1 listeners; Cooke et al., 2008)

and novel research questions (e.g., the effects of lexical difficulty on subjective effort for L1 and LX listeners).

Transparency and Openness

All data, code for analysis, and stimulus materials for both experiments are available at https://osf.io/tvwur/. The preregistration document is available at https://osf.io/etpnu. We followed Journal Article Reporting Standards (Kazak, 2018). All data were analyzed in R Version 4.2.2 (R Core Team, 2022). Data were cleaned using the *tidyverse* suite of packages (Version 2.0.0; Wickham et al., 2019), and statistical analyses were performed using the lme4 package (Version 1.1.31; Bates et al., 2015).

Experiment 1: Words

Method

Participants

To obtain our preregistered sample size of 100 participants in each of the two groups, we collected data from 220 participants, recruited online via Prolific (https://www.prolific.co). Four (three LX, one L1) were replaced because their language background on Prolific did not match what they self-reported in the demographic questionnaire. An additional participant (from the L1 group) was excluded because of low accuracy in at least one noise or difficulty condition (see preregistration for details). Four participants (all from the LX group) were replaced because they responded to 220 or fewer of the 240 trials. This was not a preregistered exclusion criterion, but we opted to replace these participants before analyzing the data because it was unclear whether the missing responses were due to a technical issue (e.g., audio files were not presented) or because those were trials that the participants skipped because they had no reasonable guess. Thus, it was not clear whether those observations should be counted as incorrect or excluded from the analysis, so we opted to replace those participants entirely. We removed the final six L1 and five LX listeners to reach our preregistered sample size of 100 L1 and 100 LX participants (V. A. Brown & Strand, 2023). Participants received $7 for 35 min of participation. Carleton College's Institutional Review Board approved all research procedures. Data collection occurred on August 16 and August 23, 2022.

All participants were 18–55 years old (L1: $M = 35.4$, $SD = 9.9$; LX: $M = 27.2$, $SD = 7.6$). L1 participants were limited to Prolific users with American IP addresses[4] who had previously specified with Prolific that English was the first language they learned. The L1 group self-reported the following demographics: 8% Asian, 79% White, 1% Native Hawaiian/other Pacific Islander, 15% Black/African American; 56% male, 39% female; 9% Hispanic/Latino, 91% not Hispanic/Latino. The LX group was limited to those who reported any language other than English as their first language on their Prolific profile, but was not geographically constrained. Participants from the LX group self-reported the following demographics: 3% Asian, 87% White, 3% American Indian/Alaska Native, 1% Black/African American; 69% male, 25% female; 22% Hispanic/Latino, 75% not Hispanic/Latino (numbers for both groups may not add to 100% because participants were permitted to skip questions or select multiple options).

As expected, L1 listeners had higher mean ratings of self-reported proficiency at understanding English (M: 9.9 vs. 8.2 out of 10) and current exposure to English with friends (M: 9.7 vs. 4.0 out of 10). L1 participants also began learning English at a younger age than LX participants (M: 1.0 vs. 7.8 years). LX listeners in this sample had relatively high self-reported proficiency at speaking, understanding, reading, and writing (means ranging from 7.02 to 9.03 on a 10-point scale), though in all cases self-reported proficiency was lower for the LX group than the L1 group. For the majority of LX listeners, English was the second language learned and the second most dominant language. See Supplemental Materials for more information about participant demographics.

Stimuli

Stimuli consisted of a subset of consonant–vowel–consonant words from a phonetically transcribed dictionary (Balota et al., 2007). The database contains values for the number of phonological neighbors, average neighborhood frequency, and lexical frequency (Brysbaert & New, 2009). We selected lexically easy words by identifying words that had frequencies above the mean for all consonant–vowel–consonant words, neighborhood densities below the mean, and neighborhood frequencies below the mean. Lexically hard words were below the mean in frequency and above the mean in neighborhood density and neighborhood frequency. Those criteria generated more candidate words than necessary for this experiment, so we selected 120 easy and 120 hard words that were the most extreme in the appropriate direction on all three values. Stimuli were recorded by a female native English speaker without a strong regional dialect using a Blue Yeti microphone in Audacity (Version 2.4.2) and equated for root-mean-square amplitude using Adobe Audition (Version 22.5.0.51).

Masking noise consisted of steady-state speech-shaped noise generated in Praat (see Winn, 2018, for script for generating the noise) that matched the long-term average spectrum of the word stimuli. In the condition with noise, the masker was presented at a signal-to-noise ratio of 5 dB. This background noise level was intended to be challenging but still render performance levels above the floor, even for lexically hard words, and was chosen via informal piloting by the research team.

Procedure

The experiment was programmed and distributed using Gorilla Experiment Builder (Anwyl-Irvine et al., 2020). Participants were instructed to wear headphones and complete the task in a quiet space. Before beginning the study, participants completed a sound check in order to ensure they could comfortably hear the audio and were told not to adjust volume levels after the study started. Next, participants completed a headphone screening for web-based auditory experiments in which they were required to identify which of three 200-Hz sinusoidal tones was the quietest, a task that cannot be reliably completed with speakers (see Woods et al., 2017, for details). The screening consisted of six trials, and participants could only complete the study if they correctly identified the quiet tone

[4] This criterion was applied to mimic typical inclusion criteria in speech research.

in all six trials. If they failed the headphone screener on the first attempt, they had the opportunity to repeat it one additional time.

The word task was divided into 16 blocks that contained 15 words each (240 words total, 60 in each condition). Within each block, all words were either lexically hard or lexically easy, and each block was presented with or without masking noise. The block composition was counterbalanced so every word was presented either with or without background noise across participants, but each participant heard each word only once. The order of the blocks was randomized. After each word, participants were asked to type what they heard into a text box, guessing when unsure. The next word was presented after a variable interstimulus interval of 500–1,500 ms (in increments of 250 ms). For blocks with noise, the noise started at the onset of the trial, and the words were presented 250–500 ms later (in increments of 250 ms).

After each block, participants were asked to rate their perceived effort by completing four of the six questions from the NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988). We omitted the two questions related to temporal and physical demand due to lack of relevance to the speech identification task. An unnumbered scale with a slider was presented on the same screen for each NASA-TLX question, and the participants moved the slider to characterize their experience (see task descriptions below). The questions were listed in a consistent order (mental demand, effort, performance, and frustration) following each block.

After completing the experiment, participants filled out a questionnaire about demographics and language history (see Supplemental Materials) that was adapted from the Language Proficiency and Experiment Questionnaire (Marian et al., 2007).

Results

Accuracy was scored as either correct or incorrect at the word level. Given that participants provided typed responses, which often contain typos and misspellings, we implemented a flexible error-correcting method called Ponto (Kessler, 2017; see Supplemental Materials for more details), which gives credit for phonologically reasonable responses. Implementing flexible scoring is helpful for distinguishing errors of perception from errors of spelling (e.g., responding to target word "doubt" with "dowt"). Using a strict scoring scheme that confounds those two errors may artificially inflate differences between L1 and LX listeners because LX listeners may be more likely to spell words phonetically.

Intelligibility data were analyzed using generalized linear mixed effects models assuming a binomial distribution with a logit link function. Subjective listening effort data (NASA-TLX; Hart & Staveland, 1988) were analyzed using mixed effects models assuming a Gaussian distribution with an identity link function. Statistical significance was evaluated by comparing nested models differing only in the effect of interest via likelihood ratio tests. Coefficient estimates are provided for effects that were significant according to the likelihood ratio test. To avoid interpreting lower-order terms in the presence of an interaction, all reported coefficients were derived from a model that did not have higher order effects. For example, a coefficient estimate for a significant two-way interaction comes from a model including all two-way interactions and all relevant simple effects, but no three-way interaction, and a coefficient estimate for a significant main effect comes from a model including the effect of interest (e.g., noise level) and all other relevant main effects (e.g., lexical difficulty), but no interactions.

We first analyzed data for the L1 and LX participants separately and then analyzed the combined L1 and LX data. For models assessing identification accuracy (as opposed to subjective effort), random effects included: random intercepts for participants and words and by-participant slopes for noise and lexical difficulty. Note that we did not model by-word random slopes for lexical difficulty because it is a feature of individual words and therefore is not manipulated within items. For models assessing subjective effort, random effects included random intercepts for participants and by-participant random slopes for noise and lexical difficulty. We did not include item random effects in any models assessing subjective listening effort because NASA-TLX scores reflect responses to blocks of trials rather than individual items. All models implemented a dummy coding scheme in which the noise, lexically easy, high constraint, and L1 conditions were coded as 0.

Identification Accuracy

The flexible scoring method (Ponto; Kessler, 2017) corrected 4.9% of trials (see R script for details regarding implementation). The results of the model comparisons are shown in Tables 1 and 2, and coefficient estimates for each model are reported below in the text. Identification accuracy data are shown in Figure 1.

L1. Word identification accuracy was higher in the quiet condition than in noise ($B = 2.34$, $SE = .14$, $z = 16.18$, $p < .001$) and higher in the lexically easy condition than the lexically difficult one ($B = -1.64$, $SE = .22$, $z = -7.37$, $p < .001$). The interaction between noise and lexical difficulty was not significant.

LX. The pattern of results for LX listeners was identical to that for L1 listeners. Word identification accuracy was higher in the quiet condition than in noise ($B = 1.51$, $SE = .10$, $z = 15.36$, $p < .001$) and higher in the lexically easy condition than the lexically difficulty one ($B = -2.02$, $SE = .17$, $z = -12.15$, $p < .001$). The interaction between noise and lexical difficulty was not significant.

L1 and LX Combined. In the next set of analyses, we assessed whether language background affects speech intelligibility and whether it moderates the effects of noise, lexical difficulty, and their interaction. Relative to L1 participants, LX participants had lower overall accuracy ($B = -1.56$, $SE = .14$, $z = -11.28$, $p < .001$), were more impaired by the presence of masking noise[5] ($B = -.60$, $SE = .08$, $z = -7.27$, $p < .001$), and showed larger effects of lexical difficulty ($B = -.45$, $SE = .09$, $z = -5.19$, $p < .001$). Including the three-way interaction among noise, language group, and lexical difficulty did not improve the fit of the model, meaning we did not find evidence that the magnitude of the noise-by-lexical difficulty interaction differed for L1 and LX participants.

Listening Effort

These analyses mirrored those described above, but the outcome of interest was NASA-TLX effort scores. Although participants

[5] Note that although the raw difference between the quiet and noise conditions was larger for the LX than the L1 group, the difference was small, and the proportional improvement going from noise to quiet was actually *smaller* in the LX group than the L1 group (see Rohrer & Arslan, 2021, for an excellent discussion of the issues associated with interpreting interactions in logistic regression models).

Table 1
Results of Model Comparisons Testing the Effects of Noise, Lexical Difficulty, and Their Interaction on Word Identification Accuracy and Subjective Listening Effort for the L1 and LX Groups Separately (Experiment 1)

Effect tested	Larger model	Group	Identification accuracy	Listening effort
Main effect of noise	*Noise* + lexical_difficulty	L1	$\chi^2_1 = 179.03$, $p < .001$	$\chi^2_1 = 113.60$, $p < .001$
		LX	$\chi^2_1 = 166.99$, $p < .001$	$\chi^2_1 = 88.64$, $p < .001$
Main effect of lexical_difficulty	Noise + *lexical_difficulty*	L1	$\chi^2_1 = 49.87$, $p < .001$	$\chi^2_1 = 48.58$, $p < .001$
		LX	$\chi^2_1 = 119.91$, $p < .001$	$\chi^2_1 = 63.47$, $p < .001$
Interaction between noise and lexical difficulty	Noise + lexical_difficulty + *noise: lexical_difficulty*	L1	$\chi^2_1 = 1.49$, $p = .22$	$\chi^2_1 = 6.22$, $p = .01$
		LX	$\chi^2_1 = 1.70$, $p = .19$	$\chi^2_1 = 49.25$, $p < .001$

Note. Italicized term is omitted in the reduced model. L1 = native; LX = non-native.

answered multiple NASA-TLX questions after each block (e.g., effort, performance), data analysis was only performed on the effort question. The other questions were included so that participants could differentiate between performance on a task and the effort they exerted to attain that level of performance (Strand et al., 2018). Participants indicated their responses to the NASA-TLX questions using an unnumbered slider, and values were transformed to a 1–100 scale for analysis. Listening effort data are shown in Figure 2.

L1. L1 participants reported significantly more effort in the noise than the quiet condition ($B = -27.28$, $SE = 1.89$, $t = -14.47$, $p < .001$) and more effort in the lexically difficult condition than the lexically easy one ($B = 6.44$, $SE = .82$, $t = 7.87$, $p < .001$). There was a significant interaction between noise and lexical difficulty indicating that the effect of lexical difficulty was more pronounced in the quiet condition than the noise condition ($B = 3.35$, $SE = 1.34$, $t = 2.50$, $p = .01$).

LX. The results for LX participants mirror those of L1 participants. LX participants reported more effort in noise than in quiet ($B = -19.55$, $SE = 1.65$, $t = -11.89$, $p < .001$) and more effort in the lexically difficult condition than the lexically easy one ($B = 9.77$, $SE = 1.04$, $t = 9.37$, $p < .001$). There was a significant interaction between noise and lexical difficulty indicating that the effect of lexical difficulty was more pronounced in quiet than in noise ($B = 10.00$, $SE = 1.41$, $t = 7.08$, $p < .001$).

L1 and LX Combined. Relative to L1 participants, LX participants reported more subjective effort overall ($B = 11.37$, $SE = 2.84$, $t = 4.01$, $p < .001$), were less affected by the presence of masking noise ($B = 7.70$, $SE = 2.50$, $t = 3.08$, $p = .002$), and showed larger effects of lexical difficulty ($B = 3.41$, $SE = 1.32$, $t = 2.58$, $p = .01$). The three-way interaction between noise, language group, and lexical difficulty was significant such that the magnitude of the noise-by-lexical difficulty interaction—which in both groups suggested that the effect of lexical difficulty was more pronounced in quiet than in noise—was stronger in the LX group than the L1 group ($B = 6.56$, $SE = 1.95$, $t = 3.37$, $p < .001$).

Discussion

The results of Experiment 1 replicate previous work demonstrating differences between L1 and LX listeners in some aspects of speech processing. LX listeners showed lower identification accuracy overall (Black & Hast, 1962; see Scharenborg & van Os, 2019, for a review) and reported more subjective effort (Peng & Wang, 2019) than L1 listeners. In addition, spoken word identification accuracy was more impaired by the presence of masking noise in LX than L1 listeners (Gat & Keith, 1978; see Lecumberri et al., 2010, for a review), and LX participants showed larger effects of lexical difficulty on speech identification accuracy than L1 participants (replicating Bradlow & Pisoni, 1999; see also Imai et al., 2005).

To our knowledge, this is the first demonstration that the participant group-by-lexical difficulty interaction is also present for listening effort such that LX listeners show larger effects of lexical difficulty than L1 listeners. This finding is consistent with the idea that LX listeners have particular difficulty identifying words that require making fine-grained phonetic distinctions (Bradlow & Alexander, 2007; Bradlow & Bent, 2002). Finally, we found evidence that LX listeners were *less* affected by background noise than L1 listeners in terms of subjective effort; this finding is

Table 2
Results of Model Comparisons for Word Identification Accuracy and Listening Effort, L1 and LX Data Combined (Experiment 1)

Effect tested	Larger model	Identification accuracy	Listening effort
Main effect of language group	Noise + lexical difficulty + *group*	$\chi^2_1 = 98.26$, $p < .001$	$\chi^2_1 = 14.10$, $p < .001$
Interaction between noise and language group	Noise + lexical difficulty + group + noise: lexical difficulty + *noise:group* + lexical difficulty:group	$\chi^2_1 = 49.99$, $p < .001$	$\chi^2_1 = 9.34$, $p = .002$
Interaction between lexical difficulty and language group	Noise + lexical difficulty + group + noise: lexical difficulty + noise:group + *lexical difficulty:group*	$\chi^2_1 = 25.08$, $p < .001$	$\chi^2_1 = 6.63$, $p = .01$
Interaction among noise, lexical difficulty, and language group	Noise + lexical difficulty + group + noise: lexical difficulty + noise:group + lexical difficulty:group + *noise:lexical difficulty: group*	$\chi^2_1 = 1.88$, $p = .17$	$\chi^2_1 = 11.31$, $p < .001$

Note. Italicized term is omitted in the reduced model. L1 = native; LX = nonnative.

unexpected and inconsistent with the results of the intelligibility analysis.

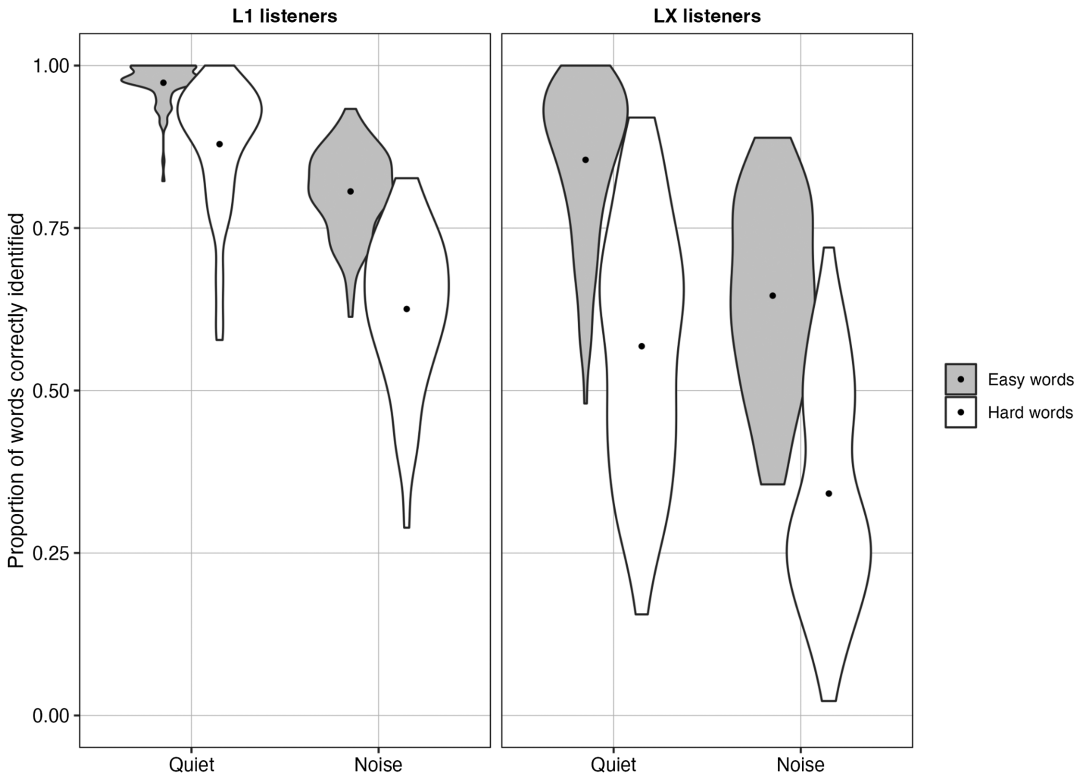
Although there were some differences in the magnitudes of the effects across groups, within the L1 and LX groups, the patterns of results were identical: (a) Both groups were impaired by noise and lexical difficulty in terms of both intelligibility and effort, (b) there was no interaction between noise and lexical difficulty on identification accuracy for either group, and (c) both groups showed a noise-by-lexical difficulty interaction on subjective effort such that the effect of lexical difficulty was more pronounced in quiet than in noise. The direction of this interaction is somewhat surprising; indeed, it might be expected that the effects of lexical difficulty would be less pronounced in quiet because the task is so easy that little effort needs to be expended regardless of lexical difficulty. However, although the finding is somewhat unexpected, the interaction was significant and in the same direction within the two groups. Thus, for the effects observed here, including both L1 and LX listeners in a sample would lead to the same general conclusions.

It is also worth noting that for each of these significant effects (noise, lexical difficulty, and the interaction for the effort data), the magnitude was larger in LX participants than L1 participants. However, particularly for the intelligibility data, there was more variability across LX listeners than L1 listeners (see Figure 1). This is unsurprising given the heterogeneity of the LX listeners. Indeed, although language background is a continuous and multidimensional variable (Luk, 2023; Weissler et al., 2023), making it dichotomous (i.e., "native" vs. "nonnative") necessarily places tighter constraints on the L1 group for three reasons: (a) Native English speakers all have the same L1; (b) given that we limited our sample to participants with IP addresses in the United States, these participants are likely to be exposed to English regularly; and (c) as L1 English speakers located in the United States, these participants likely have high English proficiency. In contrast, the LX sample differs in L1, current exposure to English (given that we did not limit our sample by country of residence), and proficiency with English.

The greater variability observed in the LX group may be driven by the phonetic properties of the L1s in that group, proficiency with English, or other factors. For example, it may be that lexical difficulty effects are stronger in individuals whose first language shares greater phonetic overlap with English (see Bradlow et al., 2010 for a discussion about classifying phonetic similarity across languages), or perhaps these effects are stronger in less proficient speakers (see Bradlow & Pisoni, 1999 for discussion of the challenges of nonnative word recognition that requires fine phonetic discrimination at the segmental level). These questions are beyond the scope of this work, but future researchers might consider exploring the extent to which these factors affect processing in LX listeners' nonnative language. In the context of this study, the key point related to differences in variability of responses across groups is that the effects of interest—noise,

Figure 1

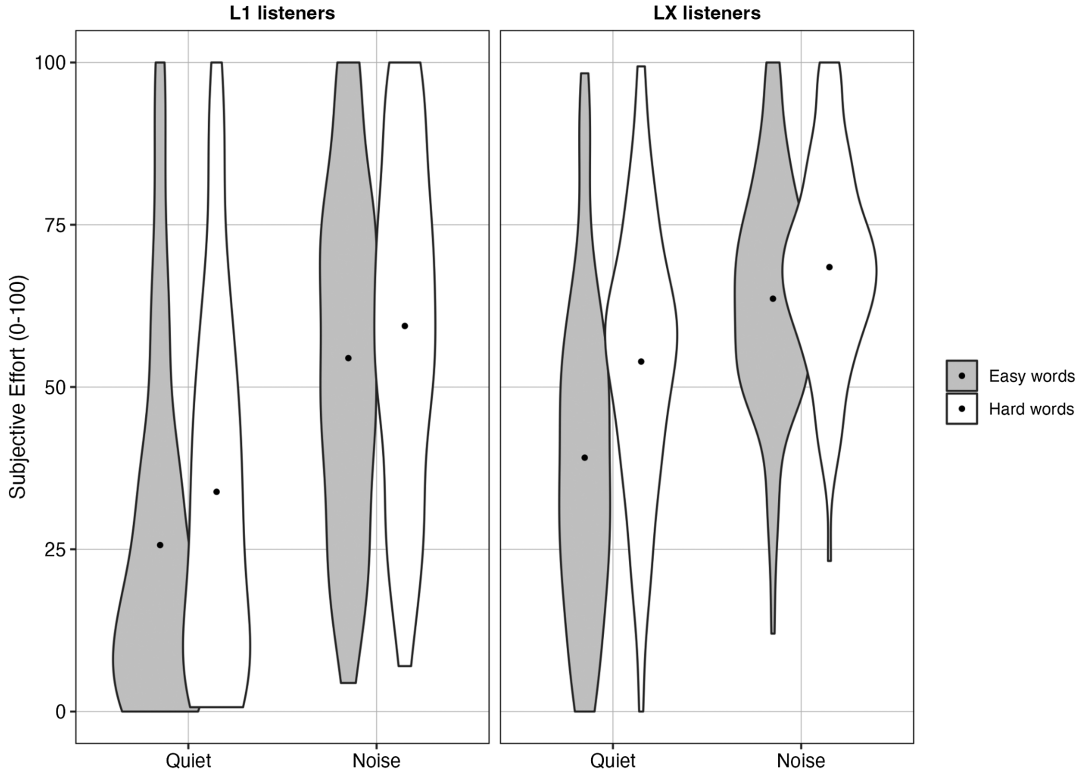
By-Participant Identification Accuracies Grouped by Listener Group, Noise, and Lexical Difficulty



Note. Dots indicate mean values for each condition. L1 = learned English before another language; LX = learned English after another language.

Figure 2

By-Participant Self-Reported Listening Effort, Grouped by Listener Group, Noise, and Lexical Difficulty



Note. Dots indicate mean values for each condition. L1 = learned English before another language; LX = learned English after another language.

lexical difficulty, and the interaction—were consistent in the two groups despite the increased variability across participants in the LX group.

Experiment 2: Sentences

Method

Experiment 2 was identical to Experiment 1 except that participants completed a sentence identification task rather than a word identification task.

Participants

In order to reach our preregistered sample size of 150 participants in each of the two groups, we collected data from 319 online participants via Prolific (https://www.prolific.co). Seven participants (all LX listeners) were replaced because their language background on Prolific did not match what they self-reported in the demographic questionnaire. An additional eight (four each from the L1 group and LX group) were excluded because of low accuracy in either noise or semantic constraint condition (see preregistration for details). We removed the final four L1 listeners to reach our preregistered sample size of 150 L1 and 150 LX participants. Participants from both the L1 and LX groups were between the ages of 18 and 55 years old (L1: $M = 35.0$, $SD = 9.1$; LX: $M = 29.1$, $SD = 8.0$). L1 participants self-reported the following demographics: 7% Asian, 77% White, 2% American Indian/Alaska Native, 13% Black/African American; 62% male, 35% female; 11% Hispanic/Latino, 85% not Hispanic/Latino. The LX group was 7% Asian, 85% White, 2% American Indian/Alaska Native, 1% Black/African American; 46% male, 50% female; 25% Hispanic/Latino, 74% not Hispanic/Latino (numbers for both participant groups may not add to 100% because participants were permitted to skip or select multiple options to the questions).

L1 speakers again had higher mean ratings of self-reported proficiency at understanding English (9.9 vs. 8.4 out of 10) and greater current exposure to English with friends (9.9 vs. 5.5 out of 10). L1 participants also began learning English at a younger age than LX participants (1.1 vs. 7.6 years). See Supplemental Materials for more information. Participants received $7 for 35 min of participation. Data collection occurred between September 6 and 15, 2022.

Stimuli

We selected a subset of 192 sentences from the Speech Perception in Noise test (Bilger et al., 1984), which consisted of 96 target words (e.g., "growl") embedded in both a high semantic constraint sentence (e.g., "The watchdog gave a warning *growl*") and a low semantic constraint one ("I had not thought about the *growl*"). Recordings were obtained from the original Speech Perception in

Noise publication (Bilger et al., 1984) and consisted of a male voice without a strong regional dialect.

Masking noise was the same as in Experiment 1, but was presented at a slightly more difficult signal-to-noise ratio of −1 dB. This change was necessary to ensure that performance was below the ceiling for semantically constrained sentences.

Procedure

The procedure for this experiment followed the conventions described in Experiment 1, including the headphone screening and questionnaires. Each participant identified 96 sentences with unique final words. Across participants, every word appeared in both constraint conditions and in both noise conditions. For each participant, the 96 target sentences were divided into 16 blocks of six sentences (four per constraint-by-noise condition), and the order in which the blocks were presented was randomized. Within each block, all sentences were either low or high constraint and were presented in quiet or in noise. Across participants, every sentence was presented in quiet and in noise approximately the same number of times. After each trial, participants were asked to identify the sentence by typing their response in a textbox. Note that although participants typed the entire sentence, we only scored accuracy for the final word.

Results

During data analysis, we realized that one target word was pluralized in the semantically constrained sentences ("fans") and singular in the semantically unconstrained sentences ("fan"), so that item was omitted from analyses.

Identification Accuracy

The conventions of the analyses for Experiment 2 mirror those of Experiment 1. The outcome variable was final word identification accuracy, and the fixed effects of interest were background noise and semantic constraint (rather than lexical difficulty). Participants and final words were included as random effects. We modeled random slopes for both noise and constraint for both random effects.[6] Ponto's flexible scoring scheme (Kessler, 2017) corrected 1.3% of responses.

The results of the model comparisons testing fixed effects of interest are shown in Tables 3 and 4, and coefficient estimates for each model are reported below in the text. Identification accuracy data are shown in Figure 3.

L1. Final word identification accuracy was higher in the quiet condition than in noise ($B = 3.02$, $SE = .26$, $z = 11.47$, $p < .001$) and higher in the constrained condition than the unconstrained one ($B = -2.23$, $SE = .20$, $z = -10.89$, $p < .001$). There was also a significant interaction between noise and constraint indicating that the effect of constraint was more pronounced in noise than in quiet ($B = 1.54$, $SE = .41$, $z = 3.76$, $p < .001$).

LX. As with the L1 group, final word identification accuracy was higher in the quiet condition than in noise ($B = 3.12$, $SE = .26$, $z = 12.01$, $p < .001$) and higher in the semantically constrained condition than the semantically unconstrained one ($B = -1.41$, $SE = .18$, $z = -7.68$, $p < .001$). As in the L1 group, the interaction between noise and constraint was significant ($B = 0.82$, $SE = .37$, $z = 2.19$, $p = .03$), with the effect of constraint being more pronounced in noise than in quiet.

L1 and LX Combined. Identification accuracy was higher for the L1 than the LX group ($B = -1.23$, $SE = .13$, $z = -9.49$, $p < .001$). The interaction between constraint and language group was significant, indicating that the effect of semantic constraint was more pronounced for the L1 group than the LX group ($B = .99$, $SE = .11$, $z = 9.11$, $p < .001$). The interaction between noise and language group was not significant, suggesting that the groups were similarly impaired by the presence of masking noise. Finally, the three-way interaction between noise, semantic constraint, and participant group was not significant, indicating that the magnitude of the interaction between noise and constraint was similar for L1 and LX participants.

Listening Effort

L1. L1 participants reported significantly more effort in the noise than the quiet condition ($B = -46.59$, $SE = 2.04$, $t = -22.82$, $p < .001$) and more effort in the unconstrained than the constrained condition ($B = 6.37$, $SE = .70$, $t = 9.11$, $p < .001$). There was a significant interaction between noise and constraint, indicating that the effect of constraint was more pronounced in noise than in quiet ($B = -3.10$, $SE = 1.15$, $t = -2.69$, $p = .007$; see Figure 4).

LX. As with the L1 participants, LX participants reported significantly more effort in the noise than the quiet condition ($B = -35.45$, $SE = 1.88$, $t = -18.90$, $p < .001$) and more effort in the unconstrained condition than the constrained one ($B = 2.06$, $SE = 0.61$, $t = 3.36$, $p < .001$). However, unlike the L1 group, interaction between noise and constraint was not significant.

L1 and LX Combined. Relative to L1 participants, LX participants reported more subjective effort overall ($B = 5.33$, $SE = 1.87$, $t = 2.85$, $p = .005$), were less affected by the presence of masking noise ($B = 11.21$, $SE = 2.77$, $t = 4.04$, $p < .001$), and showed smaller effects of constraint ($B = -4.32$, $SE = 0.93$, $t = -4.64$, $p < .001$). The three-way interaction between noise, language group, and constraint was not significant.

Although the three-way interaction was not significant, we conducted an exploratory analysis of the constraint-by-noise interaction on the combined L1 and LX data because this was the only finding in the article for which the L1 and LX groups rendered qualitatively different results (the interaction was significant for L1 participants but not LX participants). A model that included the noise-by-constraint interaction provided a better fit for the data than one without ($\chi^2_1 = 6.41$, $p = .01$), and—consistent with the results from the L1 group—the effect of constraint was more pronounced in noise than quiet ($B = -2.04$, $SE = 0.81$, $z = -2.53$, $p = .01$).[7]

[6] We preregistered that we would include sentences rather than final words as the item-level random effects. However, during data analysis, we realized that it would be preferable to use the final word instead which enables us to explicitly model the fact that the same word is appearing in both the low- and high-constraint sentences.

[7] Note, however, that the combined data set contains 50% L1 and 50% LX listeners, which likely represents a larger proportion of LX listeners than would appear in a nonrestricted sample for research conducted in the United States, particularly on college campuses. See the General Discussion section for more on how the proportion of LX listeners in a sample might affect study outcomes.

Table 3
Results of Model Comparisons Testing the Effects of Noise, Lexical Difficulty, and Their Interaction on Final-Word Sentence Identification Accuracy and Listening Effort for the L1 and LX Groups Separately (Experiment 2)

Effect tested	Larger model	Group	Identification accuracy	Listening effort
Main effect of noise	*Noise* + constraint	L1	$\chi^2_1 = 82.54$, $p < .001$	$\chi^2_1 = 225.45$, $p < .001$
		LX	$\chi^2_1 = 85.67$, $p < .001$	$\chi^2_1 = 183.49$, $p < .001$
Main effect of constraint	Noise + *constraint*	L1	$\chi^2_1 = 75.62$, $p < .001$	$\chi^2_1 = 66.35$, $p < .001$
		LX	$\chi^2_1 = 48.38$, $p < .001$	$\chi^2_1 = 10.94$, $p < .001$
Interaction between noise and constraint	Noise + constraint + *noise:constraint*	L1	$\chi^2_1 = 13.53$, $p < .001$	$\chi^2_1 = 7.24$, $p = .007$
		LX	$\chi^2_1 = 4.49$, $p = .03$	$\chi^2_1 = 0.76$, $p = .38$

Note. Italicized term is omitted in the reduced model. L1 = native; LX = nonnative.

Discussion

As in Experiment 1, the results of this experiment provide evidence that many robust findings in speech research are present in both L1 and LX listener groups, but the magnitudes of some of these effects differ across the groups. Importantly, both groups showed effects of noise and semantic constraint such that background noise impaired speech intelligibility and high constraint improved it. Further, for both groups, the effect of constraint on intelligibility was more pronounced in noise than in quiet. There were also similarities across groups in the subjective effort data: For both listener groups, subjective effort ratings were lower in quiet relative to noise and in high relative to low constraint. The only difference in the pattern of results within each group was that the interaction between noise and constraint was significant in the L1 group but not in the LX group for subjective effort. This interaction suggests that for L1 listeners, the effect of constraint on effort is more pronounced in noise than in quiet. This finding is consistent with the idea that when listening conditions are difficult, listeners have greater opportunity to capitalize on cues that facilitate speech processing (this is also true of audiovisual speech processing, whereby the addition of the visual signal has greater opportunity to reduce effort in difficult listening conditions; see, e.g., V. A. Brown & Strand, 2019). It is unclear why this interaction was significant for effort in the L1 group but not the LX group, but this may simply be driven by the fact that the effect of constraint was smaller for LX listeners. However, given that three-way interaction between noise, constraint, and listener group was not significant for either intelligibility or effort, this difference in significance for the interaction should be interpreted cautiously.

Consistent with previous work and with the results of Experiment 1, we found that LX listeners showed poorer speech identification accuracy overall (Black & Hast, 1962; Scharenborg & van Os, 2019) and reported more subjective listening effort (Peng & Wang, 2019) than L1 listeners. Our results also replicate previous work showing that L1 listeners gain greater intelligibility benefit from semantic contextual cues than LX listeners (Bradlow & Alexander, 2007; Mayo et al., 1997; Shi, 2010, 2014) and extend this finding to subjective listening effort. L1 listeners may be able to capitalize on contextual cues more efficiently than LX listeners because processing semantically constraining sentences involves knowledge of vocabulary and sentence structure—in addition to the ability to rapidly integrate these contextual cues with the bottom-up input to anticipate what words are likely to occur next—and LX listeners necessarily have less experience with these processes in English than L1 listeners (Shi, 2014).

Two somewhat puzzling findings emerged from this experiment, both involving the interaction between listener group and background noise. First, the interaction was not significant for the intelligibility data, which is inconsistent with the results of Experiment 1 as well as previous research (Gat & Keith, 1978; Lecumberri et al., 2010), and suggests that L1 and LX listeners are similarly affected by background noise. Second, the interaction was significant for the subjective effort data, but was in the opposite direction from what would be expected: The interaction indicated that LX listeners were less affected by background noise than L1 listeners. This effect also emerged in Experiment 1, and it is unclear why this would be the case.

General Discussion

The goal of this study was to evaluate whether several classic and novel findings in the spoken word and sentence identification literatures are robust to variations in the language backgrounds of the

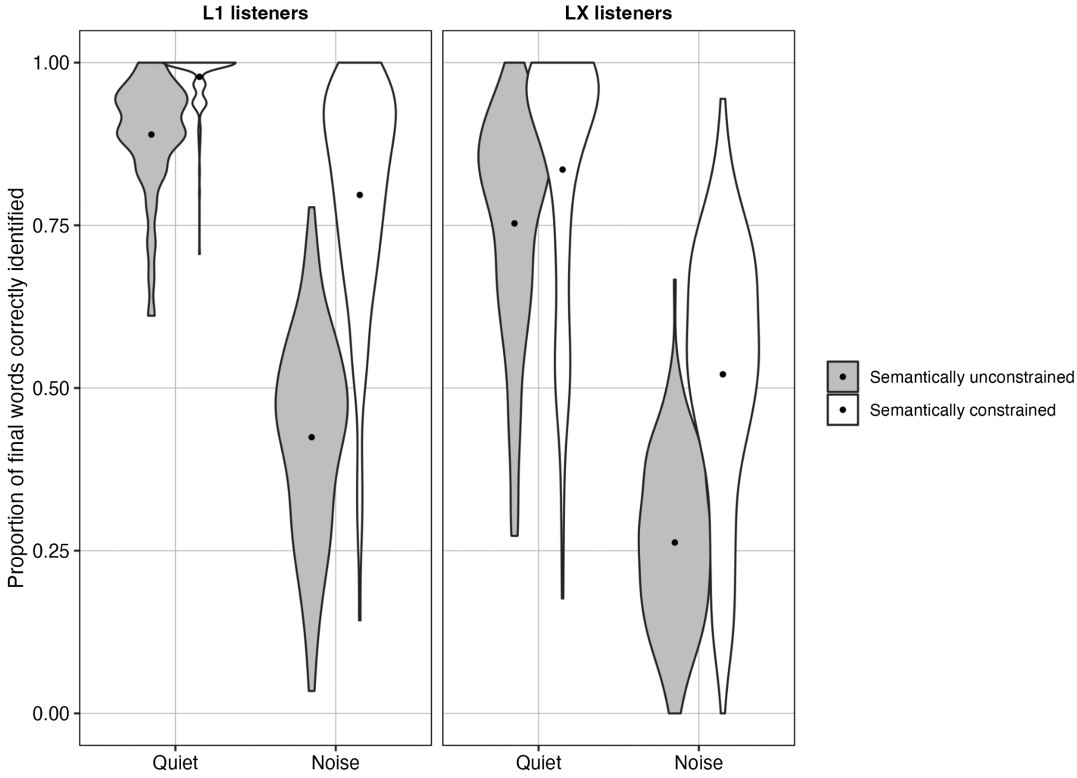
Table 4
Results of Model Comparisons for Final-Word Identification Accuracy and Listening Effort, L1 and LX Data Combined (Experiment 2)

Effect tested	Larger model	Identification accuracy	Listening effort
Main effect of language group	Noise + constraint + *group*	$\chi^2_1 = 79.08$, $p < .001$	$\chi^2_1 = 7.26$, $p = .007$
Interaction between noise and language group	Noise + constraint + group + noise:constraint + *noise:group* + constraint:group	$\chi^2_1 = 1.80$, $p = .18$	$\chi^2_1 = 16.02$, $p < .001$
Interaction between constraint and language group	Noise + constraint + group + noise:constraint + noise:group + *constraint:group*	$\chi^2_1 = 77.70$, $p < .001$	$\chi^2_1 = 20.95$, $p < .001$
Interaction among noise, semantic constraint, and language group	Noise + constraint + group + noise:constraint + noise:group + constraint:group + *noise: constraint:group*	$\chi^2_1 = 2.25$, $p = .13$	$\chi^2_1 = 1.77$, $p = .18$

Note. Italicized term is omitted in the reduced model. L1 = native; LX = nonnative.

Figure 3

By-Participant Final-Word Identification Accuracies, Grouped by Listener Group, Noise, and Semantic Constraint



Note. Dots indicate mean values for each condition. L1 = learned English before another language; LX = learned English after another language.

participants. One key takeaway is that across two experiments, all of the main effects we tested—including effects of background noise, lexical difficulty, and semantic cues on both speech identification accuracy and subjective listening effort—were significant and in the same direction in both groups. Indeed, regardless of whether we analyzed data exclusively from L1 or LX listeners, words were identified less accurately when they were lexically difficult and when they were presented in background noise, and sentences were identified less accurately when they were less predictable and when they were presented in background noise. Further, three of the four within-group interactions (noise-by-lexical difficulty for both identification accuracy and subjective effort and noise-by-constraint for identification accuracy) were significant and in the same direction for L1 and LX listeners. Thus, for most of the within-group analyses, the conclusions would have been identical regardless of whether the sample consisted of L1 or LX listeners (the only exception was the noise-by-constraint interaction for effort; more on this below).

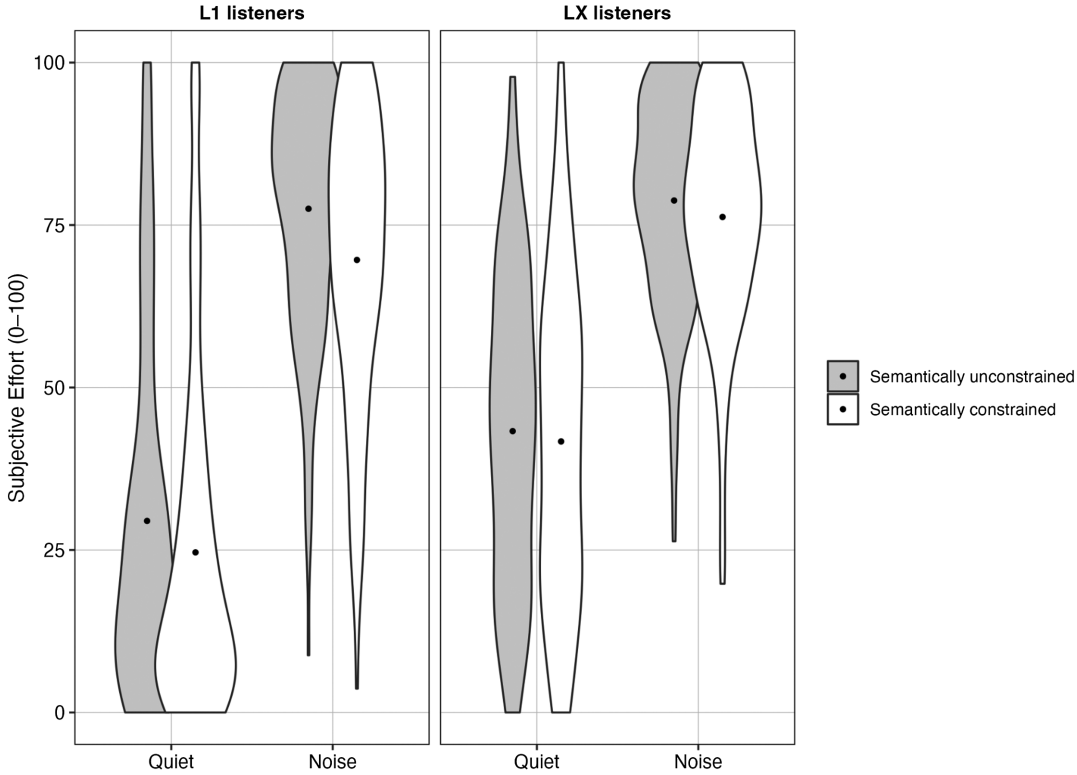
Although the patterns of results within the two groups were mostly consistent, we found that seven of the eight interactions between the variable of interest (i.e., noise, lexical difficulty, or constraint) and language group were significant, suggesting that in most cases, the magnitudes of the effects differed for L1 and LX listeners. For example, consistent with previous work, these analyses revealed that LX listeners were more affected by lexical difficulty (Bradlow & Pisoni, 1999) and were less affected by semantic constraint (Bradlow & Alexander, 2007) relative to L1 listeners.

Looking only at the main effects and within-group interactions, one could conclude that—for the tasks studied here—including LX participants in the sample would be unlikely to change study outcomes. In contrast, given the interactions between language group and the variables of interest, one could also conclude that because L1 and LX participants are differently affected by some of the manipulations of interest, including LX listeners in the sample may change outcomes. How can we reconcile these two opposing conclusions? Below, we outline recommendations for language researchers that take into account the goals of the researcher and the nature of the variables and outcomes of interest. We begin with scenarios in which it may be informative and necessary to account for the language backgrounds of the participants or possibly exclude LX listeners from the sample, and end with situations in which including LX listeners in the sample is unlikely to affect the pattern of results.

A scenario in which it can unequivocally be argued that it would be necessary to collect data from both L1 and LX listeners and analyze the groups separately (and/or test interactions between language background and other variables) is if the researcher is specifically interested in how language background affects study outcomes and whether other effects of interest are moderated by language background. For example, if the goal of the research is to

Figure 4

By-Participant Self-Reported Listening Effort, Grouped by Listener Group, Noise, and Semantic Constraint



Note. Dots indicate mean values for each condition. L1 = learned English before another language; LX = learned English after another language.

evaluate whether LX listeners gain similar intelligibility benefits from seeing the talker relative to L1 listeners (e.g., Xie et al., 2014), it would be necessary to collect data from both listener groups and test the interaction between modality and group (and perhaps test the effect of modality within the two groups separately if appropriate).

The previous example describes a situation in which careful attention must be paid to the language background of the participants because group differences are central to the research question; that is, the sample must contain data from both L1 and LX listeners given the design of the study. But what course of action should a researcher take if group differences are *not* part of the research question? When should LX listeners be excluded from the sample? We argue that in these cases, whether or not to include LX listeners in the sample—and if these individuals are included in the sample, whether and how language background should be included in statistical analyses—depends on the particular research question and the variables and outcomes of interest.

One situation in which careful attention to the language background of the participants may be necessary even when group differences are not central to the research question per se is if language differences are known to interact with the variable of interest in predicting the outcome, *and the magnitude of the estimate is crucial to the research question*. Although both this situation and the previous one involve knowledge of potential interaction effects, group differences are of interest in the former but are an inconvenience (i.e., an extraneous variable) in the latter. As an example, this may be relevant for research on the relationship between voice onset time and voicing perception: Voice onset time distributions are known to differ across speakers of different languages (Flege & Eefting, 1987), so whether particular syllable tokens are perceived as voiced or voiceless will differ for speakers of English and Spanish. Thus, if the goal of the study is to obtain an accurate estimate of the relationship between voice onset time and voicing perception, including LX listeners in the sample but not the analyses will lead to biased estimates that do not accurately reflect the relationship between voice onset time and voicing perception in either group. In this case, the researcher may choose to limit their sample to L1 listeners (sometimes referred to as "control by elimination"; Chen & Krauss, 2005), or, if appropriate, incorporate language differences into their research question by testing the interaction between voice onset time and language background (sometimes referred to as "control by inclusion") and possibly analyzing the data in the two groups separately.

Note that simply including language background as a covariate in the example above would not make the estimate for the effect of voice onset time less biased: Voice onset time was experimentally manipulated and is therefore uncorrelated with language background, and this lack of covariance between the predictors means that the coefficient estimates will be the same[8] regardless of whether

[8] Or nearly the same, depending on whether the design is fully balanced and the predictors are completely uncorrelated.

the other variable is included in the model. However, in some situations, language background is correlated with both the predictor and the outcome of interest. For example, consider a researcher who is interested in how individual differences in category boundaries between voiced and voiceless consonants affect voicing perception. Given that language background affects category boundaries and therefore voicing perception, including language background as a covariate would enable researchers to evaluate the unique influence of category boundaries above and beyond the influence of language background.[9] Thus, in cases like these, we suggest that researchers either limit their sample to L1 listeners or preferably include LX listeners in their sample and model language background as a covariate in their statistical analyses.

Finally, our results also illustrate that there may be situations in which collecting data from LX listeners is justifiable even without accounting for differences in language background in statistical analyses. In our case, for example, if the goal of our research had been to simply evaluate whether lexical difficulty effects are more pronounced in easy levels of background noise than in quiet, we would have arrived at the same conclusion regardless of our sample. At this point, it is important to note that in this study we included separate groups consisting of either 100% L1 or 100% LX listeners, and conclusions were identical in 11 of the 12 analyses despite the extreme groups design. In reality, the proportion of LX listeners in a typical sample would almost certainly be much lower (e.g., when drawing from undergraduates on a college campus). For the finding that would have led to different conclusions in the L1 and LX groups (a significant interaction between noise and constraint in the L1 but not the LX group for subjective effort), the effects were in the same direction in the two groups and indeed still emerged in the combined data with 50% L1 and 50% LX listeners. Thus, the interaction would also likely emerge in a typical undergraduate sample that consists of mostly L1 listeners. Relatedly, the LX listeners in our study were not geographically constrained and therefore consisted of individuals living around the world with varying levels of experience with and exposure to American English. Thus, the magnitudes of the L1/LX differences we observed—as well as the variability of responses within the LX group—are likely larger than what may be observed in studies conducted on American college campuses, even if those studies include LX listeners.

Although it may seem counterintuitive not to account for a variable that is known to affect study outcomes, this is common practice for individual differences variables that are known to be related to speech processing. For example, a substantial body of literature shows that individual differences in working memory capacity predict a host of outcomes related to language processing: Individuals with higher working memory capacity tend to retain more information from spoken passages (Daneman & Merikle, 1996), be quicker to anticipate upcoming words based on grammatical cues (Huettig & Janse, 2016), and be better lipreaders (Feld & Sommers, 2009). In fact, there is some evidence that people with better working memory capacity show greater facilitation from semantic cues (Zekveld et al., 2013). However, it is not common practice to only collect speech intelligibility data from participants with high working memory capacity nor to include working memory capacity in statistical analyses when the research question is unrelated to working memory. Analogously, although there are certainly differences in speech processing between L1 and LX listeners, when the research question is not strictly about differences in language background, it may not be necessary to only sample L1 listeners. In the absence of a compelling reason to limit the sample to L1 participants, there are distinct advantages to including LX listeners: Not only is this practice likely to make research findings more generalizable, but it also helps to correct historic inequities in who can participate in (and therefore benefit from) the research.

Taken together, the results of this study lead to a key takeaway: We recommend that future researchers carefully consider whether L1/LX status is likely to affect outcomes, taking into account (a) theoretical reasons that group differences might be expected, (b) whether these differences are large enough to affect study outcomes, and (c) whether the number of LX listeners in the sample is likely to be large enough to meaningfully influence the effects of interest. If a researcher concludes that LX listeners should be excluded from the sample, we suggest that they pay particular attention to how labels like "LX" or "native" are being defined in the context of the study (e.g., related to proficiency, age of acquisition, etc.; see Cheng et al., 2021). Although we tested a limited range of effects, this work represents a proof of concept: Previous research on the topics studied here that excluded LX participants may have led to similar conclusions if those participants had been included. This may be true of other fields as well. We hope that this work will encourage the broader research community to move away from the default of excluding "nonnative" participants without justification—a practice that our survey of the 100 most cited articles revealed to be quite prevalent in spoken language research—and instead make decisions about sample composition on a study-by-study basis.

Constraints on Generality

The purpose of this article was to explicitly test whether and how a commonly used constraint on participation in speech research and other cognitive domains affects study outcomes, and we conclude that researchers should think carefully about how the composition of their sample may or may not affect results. However, we did not explicitly test language proficiency of the LX participants in our analyses; given that our LX listeners had relatively high self-reported English proficiency, it may be that the results of our experiments would be different for low-proficiency English speakers. Additionally, the extent to which language background affects outcomes certainly depends on the nature of the specific tasks used. The tasks employed here were not intended to be representative of all speech perception tasks; it is certainly possible that other tasks produce larger differences in outcomes between L1 and LX groups.

[9] This is different from the previous situation because although language background affects both voice onset time distributions and voicing perception, in an experimental study testing the relationship between voice onset time (a manipulated variable) and voicing perception, language background is not related to the voice onset time variable (assuming a balanced design). Rather, it moderates the relationship between the voice onset time variable and voicing perception (i.e., it interacts with voice onset time in predicting voicing perception).

References

Abrahamsson, N., & Hyltenstam, K. (2009). Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny. *Language Learning*, *59*(2), 249–306. https://doi.org/10.1111/j.1467-9922.2009.00507.x

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bernolet, S., Collina, S., & Hartsuiker, R. J. (2016). The persistence of syntactic priming revisited. *Journal of Memory and Language*, *91*, 99–116. https://doi.org/10.1016/j.jml.2016.01.002

Berntsen, D., Hoyle, R. H., & Rubin, D. C. (2019). The Autobiographical Recollection Test (ART): A measure of individual differences in autobiographical memory. *Journal of Applied Research in Memory and Cognition*, *8*(3), 305–318. https://doi.org/10.1037/h0101839

Bilger, R. C., Nuetzel, J. M., Rabinowitz, W. M., & Rzeczkowski, C. (1984). Standardization of a test of speech perception in noise. *Journal of Speech and Hearing Research*, *27*(1), 32–48. https://doi.org/10.1044/jshr.2701.32

Black, J. W., & Hast, M. H. (1962). Speech reception with altering signal. *Journal of Speech and Hearing Research*, *5*(1), 70–75. https://doi.org/10.1044/jshr.0501.70

Borghini, G., & Hazan, V. (2018). Listening effort during sentence processing is increased for non-native listeners: A pupillometry study. *Frontiers in Neuroscience*, *12*, Article 152. https://doi.org/10.3389/fnins.2018.00152

Bradlow, A. R., & Alexander, J. A. (2007). Semantic and phonetic enhancements for speech-in-noise recognition by native and non-native listeners. *The Journal of the Acoustical Society of America*, *121*(4), 2339–2349. https://doi.org/10.1121/1.2642103

Bradlow, A. R., Clopper, C., Smiljanic, R., & Walter, M. A. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, *52*(11–12), 930–942. https://doi.org/10.1016/j.specom.2010.06.003

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America*, *112*(1), 272–284. https://doi.org/10.1121/1.1487837

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

Broersma, M., & Scharenborg, O. (2010). Native and non-native listeners' perception of English consonants in different types of noise. *Speech Communication*, *52*(11–12), 980–995. https://doi.org/10.1016/j.specom.2010.08.010

Brown, B., Tusmagambet, B., Rahming, V., Tu, C.-Y., DeSalvo, M. B., & Wiener, S. (2023). Searching for the "native" speaker: A preregistered conceptual replication and extension of Reid, Trofimovich, and O'Brien (2019). *Applied Psycholinguistics*, *44*(4), 475–494. https://doi.org/10.1017/S0142716423000127

Brown, V. A., & Strand, J. F. (2019). About face: Seeing the talker improves spoken word recognition but increases listening effort. *Journal of Cognition*, *2*(1), Article 44. https://doi.org/10.5334/joc.89

Brown, V. A., & Strand, J. F. (2023). Preregistration: Practical considerations for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research*, *66*(6), 1889–1898. https://doi.org/10.1044/2022_JSLHR-22-00317

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

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*, Article 715843. https://doi.org/10.3389/fpsyg.2021.715843

Chen, P. Y., & Krauss, A. D. (2005). Experiments, psychology. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 911–918). Elsevier. https://doi.org/10.1016/B0-12-369398-5/00327-3

Cooke, M., Garcia Lecumberri, M. L., & Barker, J. (2008). The foreign language cocktail party problem: Energetic and informational masking effects in non-native speech perception. *The Journal of the Acoustical Society of America*, *123*(1), 414–427. https://doi.org/10.1121/1.2804952

Cutler, A., Weber, A., Smits, R., & Cooper, N. (2004). Patterns of English phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, *116*(6), 3668–3678. https://doi.org/10.1121/1.1810292

Daneman, M., & Merikle, P. M. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin & Review*, *3*(4), 422–433. https://doi.org/10.3758/BF03214546

Debenport, E. (2011). As the Rez Turns: Anomalies within and beyond the boundaries of a Pueblo community. *American Indian Culture and Research Journal*, *35*(2), 87–110. https://doi.org/10.17953/aicr.35.2.e22v3341216010g

de Bruin, A., Treccani, B., & Della Sala, S. (2015). Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science*, *26*(1), 99–107. https://doi.org/10.1177/0956797614557866

Dewaele, J.-M. (2018). Why the dichotomy "L1 versus LX user" is better than "native versus non-native speaker". *Applied Linguistics*, *39*(2), 236–240. https://doi.org/10.1093/applin/amw055

Feld, J. E., & Sommers, M. S. (2009). Lipreading, processing speed, and working memory in younger and older adults. *Journal of Speech, Language, and Hearing Research*, *52*(6), 1555–1565. https://doi.org/10.1044/1092-4388(2009/08-0137)

Flege, J. E. (1992). Speech learning in a second language. In C. Ferguson, L. Menn, & C. Stoel-Gammon (Eds.), *Phonological development: Models, research* (pp. 565–604). York Press.

Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*(1), 67–83. https://doi.org/10.1016/S0095-4470(19)30538-8

Gat, I. B., & Keith, R. W. (1978). An effect of linguistic experience: Auditory word discrimination by native and non-native speakers of English. *Audiology*, *17*(4), 339–345. https://doi.org/10.3109/00206097809101303

Gilmore, A. W., Nelson, S. M., & McDermott, K. B. (2016). The contextual association network activates more for remembered than for imagined events. *Cerebral Cortex*, *26*(2), 611–617. https://doi.org/10.1093/cercor/bhu223

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Hatukai, T., & Algom, D. (2017). The Stroop incongruity effect: Congruity relationship reaches beyond the Stroop task. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1098–1114. https://doi.org/10.1037/xhp0000381

Hazan, V., & Simpson, A. (2000). The effect of cue-enhancement on consonant intelligibility in noise: Speaker and listener effects. *Language and Speech*, *43*(3), 273–294. https://doi.org/10.1177/002383090004300301

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Huettig, F., & Janse, E. (2016). Individual differences in working memory and processing speed predict anticipatory spoken language processing in

the visual world. *Language, Cognition and Neuroscience*, *31*(1), 80–93. https://doi.org/10.1080/23273798.2015.1047459

Imai, S., Walley, A. C., & Flege, J. E. (2005). Lexical frequency and neighborhood density effects on the recognition of native and Spanish-accented words by native English and Spanish listeners. *The Journal of the Acoustical Society of America*, *117*(2), 896–907. https://doi.org/10.1121/1.1823291

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kessler, B. (2017). *Ponto*. http://spell.psychology.wustl.edu/ponto/

Kutlu, E., & Hayes-Harb, R. (2023). Towards a just and equitable applied psycholinguistics. *Applied Psycholinguistics*, *44*(3), 293–300. https://doi.org/10.1017/S0142716423000280

Lecumberri, M. L. G., Cooke, M., & Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech Communication*, *52*(11–12), 864–886. https://doi.org/10.1016/j.specom.2010.08.014

Ledgerwood, A., Hudson, S. 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), 937–959. https://doi.org/10.1177/17456916211036654

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

Luk, G. (2023). Justice and equity for whom? Reframing research on the "bilingual (dis)advantage". *Applied Psycholinguistics*, *44*(3), 301–315. https://doi.org/10.1017/S0142716422000339

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50*(4), 940–967. https://doi.org/10.1044/1092-4388(2007/067)

Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of second-language acquisition and perception of speech in noise. *Journal of Speech, Language, and Hearing Research*, *40*(3), 686–693. https://doi.org/10.1044/jslhr.4003.686

Peng, Z. E., & Wang, L. M. (2019). Listening effort by native and nonnative listeners due to noise, reverberation, and talker foreign accent during English speech perception. *Journal of Speech, Language, and Hearing Research*, *62*(4), 1068–1081. https://doi.org/10.1044/2018_JSLHR-H-17-0423

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., & Wingfield, A. (2016). Hearing impairment and cognitive energy: The Framework for Understanding Effortful Listening (FUEL). *Ear and Hearing*, *37*(Suppl. 1), 5S–27S. https://doi.org/10.1097/AUD.0000000000000312

Poole, B. J., & Kane, M. J. (2009). Working-memory capacity predicts the executive control of visual search among distractors: The influences of sustained and selective attention. *Quarterly Journal of Experimental Psychology*, *62*(7), 1430–1454. https://doi.org/10.1080/17470210802479329

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of Homo sapiens: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

R Core Team. (2022). *The R project for statistical computing* (Version 4.2.2) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Roberts, S. O., & Mortenson, E. (2023). Challenging the White = neutral framework in psychology. *Perspectives on Psychological Science*, *18*(3), 597–606. https://doi.org/10.1177/17456916221077117

Rogers, C. L., Lister, J. J., Febo, D. M., Besing, J. M., & Abrams, H. B. (2006). Effects of bilingualism, noise, and reverberation on speech perception by listeners with normal hearing. *Applied Psycholinguistics*, *27*(3), 465–485. https://doi.org/10.1017/S014271640606036X

Rohrer, J. M., & Arslan, R. C. (2021). Precise answers to vague questions: Issues with interactions. *Advances in Methods and Practices in Psychological Science*, *4*(2), Article 25152459211007368. https://doi.org/10.1177/25152459211007368

Scharenborg, O., & van Os, M. (2019). Why listening in background noise is harder in a non-native language than in a native language: A review. *Speech Communication*, *108*, 53–64. https://doi.org/10.1016/j.specom.2019.03.001

Shi, L.-F. (2010). Perception of acoustically degraded sentences in bilingual listeners who differ in age of English acquisition. *Journal of Speech, Language, and Hearing Research*, *53*(4), 821–835. https://doi.org/10.1044/1092-4388(2010/09-0081)

Shi, L.-F. (2014). Measuring effectiveness of semantic cues in degraded English sentences in non-native listeners. *International Journal of Audiology*, *53*(1), 30–39. https://doi.org/10.3109/14992027.2013.825052

Stern, H. (1983). *Fundamental concepts of language teaching: Historical and interdisciplinary perspectives on applied linguistic research*. Oxford University Press.

Strand, J. F., Brown, V. A., Merchant, M. B., Brown, H. E., & Smith, J. (2018). Measuring listening effort: Convergent validity, sensitivity, and links with cognitive and personality measures. *Journal of Speech, Language, and Hearing Research*, *61*(6), 1463–1486. https://doi.org/10.1044/2018_JSLHR-H-17-0257

Syed, M. (2021, June 10). WEIRD times: Three reasons to stop using a silly acronym. *Get Syeducated*. https://getsyeducated.blogspot.com/2021/06/weird-times-three-reasons-to-stop-using.html

U.S. Census Bureau. (2020). *ACS 5-year data*. https://data.census.gov/mdat

Weissler, R. E., Drake, S., Kampf, K., Diantoro, C., Foster, K., Kirkpatrick, A., Preligera, I., Wesson, O., Wood, A., & Baese-Berk, M. M. (2023). Examining linguistic and experimenter biases through "non-native" versus "native" speech. *Applied Psycholinguistics*, *44*(4), 460–474. https://doi.org/10.1017/S0142716423000115

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), Article 1686. https://doi.org/10.21105/joss.01686

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

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics*, *79*(7), 2064–2072. https://doi.org/10.3758/s13414-017-1361-2

Xie, Z., Yi, H.-G., & Chandrasekaran, B. (2014). Nonnative audiovisual speech perception in noise: Dissociable effects of the speaker and listener. *PLOS ONE*, *9*(12), Article e114439. https://doi.org/10.1371/journal.pone.0114439

Zekveld, A. A., Rudner, M., Johnsrude, I. S., & Rönnberg, J. (2013). The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *The Journal of the Acoustical Society of America*, *134*(3), 2225–2234. https://doi.org/10.1121/1.4817926

Received July 13, 2023
Revision received May 1, 2024
Accepted June 17, 2024 ■