Tessa Bent*, Holly Lind-Combs, Rachael F. Holt and Cynthia Clopper

Perception of regional and nonnative accents: a comparison of museum laboratory and online data collection

https://doi.org/10.1515/lingvan-2021-0157
Received December 21, 2021; accepted December 13, 2022; published online May 15, 2023

Abstract: Online testing for behavioral research has become an increasingly used tool. Although more researchers have been using online data collection methods, few studies have assessed the replicability of findings for speech intelligibility tasks. Here we assess intelligibility in quiet and two noise-added conditions for several different accents of English (Midland American, Standard Southern British, Scottish, German-accented, Mandarin-accented, Japanese-accented, and Hindi-English bilingual). Participants were tested in person at a museum-based laboratory and online. Results showed little to no difference between the two settings for the easier noise condition and in quiet, but large performance differences in the most difficult noise condition with an advantage for the participants tested online. Technology-based variables did not appear to drive the setting effect, but experimenter presence may have influenced response strategy for the in-person group and differences in demographics could have provided advantages for the online group. Additional research should continue to investigate how setting, demographic factors, experimenter presence, and motivational factors interact to determine performance in speech perception experiments.

Keywords: nonnative accent; online testing; regional accent; remote data collection; word identification

1 Introduction

There has been growing interest in using online data collection for behavioral research studies, which only intensified during the COVID-19 pandemic. Many advantages have been noted for online data collection methods, including more diverse participant pools, particularly compared to university-based laboratories, lower costs to the researcher, faster and more efficient data collection, less experimenter bias or impact of participant expectations, and greater ease of sharing with other researchers (Buhrmester et al. 2011, 2018; Casler et al. 2013; Rezlescu et al. 2020; Slote and Strand 2016; Yoho et al. 2019). However, researchers have also noted some drawbacks. In some studies, participants tested online have shown lower accuracy and longer response latencies (Dandurand et al. 2008; Pare and Cree 2009) and higher dropout rates (Cronk and West 2002; Dandurand et al. 2008) than in-lab participants. Further, there is a loss of control over many variables when testing online including technology employed by participants and the presence of environmental distractions or inattention (Buhrmester et al. 2018). Researchers must take much more care in writing instructions for web-based tasks because they are not physically present to address any confusions (Ramsey et al. 2016). Finally, there have been concerns raised about the lack of naivety for the participants, particularly when Amazon Mechanical Turk is used with commonly employed tasks (Peer et al. 2017).

One of the central questions regarding these data collection methods is whether findings from laboratory-based studies can be replicated with online testing. Numerous studies have shown replication for a range of cognitive and linguistic tasks using participants tested online (Balota et al. 2001; Crump et al. 2013; Ramsey et al.

***Corresponding author: Tessa Bent**, Speech, Language and Hearing Sciences, Indiana University, Bloomington, IN, USA, E-mail: tbent@iu.edu. https://orcid.org/0000-0001-7604-1835
Holly Lind-Combs and Rachael F. Holt, Speech and Hearing Science, The Ohio State University, Columbus, OH, USA
Cynthia Clopper, Linguistics, The Ohio State University, Columbus, OH, USA

2016; Sprouse 2011). However, few studies have used online testing for auditory-based experiments and specifically word recognition experiments. Beyond the general concerns about online testing, researchers have noted additional concerns for studies that include auditory or audiovisual stimuli, such as lack of control over the headphones or speakers, output volume, and ambient noise levels (Merchant et al. 2021; Slote and Strand 2016). Despite these concerns, online testing has become more prevalent in speech perception research (Borrie et al. 2017; Burchill et al. 2018; Jiao et al. 2019; Liu and Jaeger 2018; Xie et al. 2018; Yoho et al. 2019; Yoho and Borrie 2018), including work investigating perception of nonnative speech (Cooper and Bradlow 2018; Melguy and Johnson 2021; Vaughn 2019).

Although online testing has become more common, there are few published reports assessing the replicability of online versus in-lab results for intelligibility tasks. If more researchers shift to online data collection and these data become the foundation upon which new theories are built, an understanding of if and how data collected online differ from data collected in laboratory studies will be essential. The extant studies have not been consistent in their findings regarding how online and in-lab participant performance may differ. In several studies, participants recruited and tested online show lower word recognition accuracy than participants tested in the lab for synthesized, masked, distorted, and filtered speech (Cooke et al. 2011; Cooke and Garcia Lecumberri 2021; Mayo et al. 2012; Slote and Strand 2016; Wolters et al. 2010). Other studies have shown similar word recognition performance across the two testing modalities (Lansford et al. 2016) when headphone quality is considered (Cooke and Garcia Lecumberri 2021). One study showed an intelligibility advantage for those tested online for a Spanish-accented speaker in noise compared to in-lab participants (Vaughn 2019). Although absolute intelligibility has differed between testing modalities across studies, relative intelligibility levels appear more stable. For example, intelligibility differences have been maintained across speech synthesis systems (Wolters et al. 2010), listening conditions including signal-to-noise ratios and masker types (Cooke et al. 2011), speaking styles (Mayo et al. 2012), specific lexical items (Slote and Strand 2016), and other speech intelligibility tasks (Cooke and Garcia Lecumberri 2021).

Most recent studies used Amazon Mechanical Turk (MTurk) for participant recruitment and testing. Sample sizes ranged from 40 to 260 participants per experiment with exclusion rates ranging from 5 to 65 % of recruited participants. Nearly all studies required listeners to be monolingual or native speakers of English with some also having exclusion criteria for hearing, speech, and language impairments. Although listeners tended to be young adults, most studies included listeners of any age. Differences in demographic characteristics across participant populations may account for some of the discrepancies in results described above. Two studies (Cooke and Garcia Lecumberri 2021; Cooper and Bradlow 2018) had listeners who were tested online but recruited from university student populations.

The goal of this study was to determine how intelligibility of talkers with a range of different regional and nonnative accents would differ depending on data collection setting, specifically comparing a museum-based laboratory and online testing. The examination of intelligibility across different accents provides opportunities to examine how specific phonemic and suprasegmental aspects of the speech signal may impact a listener's ability to accurately map the acoustic-phonetic signal onto words in their lexicon (e.g., Adank et al. 2009; Bent et al. 2016, 2021; Clopper and Bradlow 2008; Floccia et al. 2009). Indeed, the perception of less-familiar accents are central elements in theories and frameworks of listening effort and speech adaptation, such as the Ideal Adaptor Framework (Kleinschmidt and Jaeger 2015), the Ease of Language Understanding model (Rönnberg et al. 2013), and the Framework for Understanding Effortful Listening (Pichora-Fuller et al. 2016). Furthermore, decreases in intelligibility for speakers with nonlocal accents, particularly in difficult listening conditions (e.g., in noise), are linked with more negative language attitudes, suggesting that there may be substantial social and professional consequences of intelligibility differences across talkers (Dragojevic and Giles 2016).

Although this study builds on prior work that compared laboratory-based studies to online data collection, this comparison differs in that we consider a museum-based laboratory (Wagner et al. 2015) relative to virtual data collection using Prolific, an online participant pool specifically designed for research purposes. Additionally, no prior studies have tested whether findings regarding intelligibility for a range of accents replicates across in-person and virtual data collection modalities nor have they tested both quiet and noise-added conditions with the same stimuli.

2 Methods

2.1 Participants

The listeners included 408 adult American English–speaking monolingual listeners between the ages of 18 and 35. This age range was selected to reduce the likelihood of recruiting participants with age-related hearing loss or cognitive decline. Of these, 264 were tested in person at a science museum, and 144 were recruited through Prolific and tested online. All participants had self-reported typical hearing and language. An additional 47 participants were tested but their data were not included (museum: n = 30 or 10 %; online: n = 17 or 8 %; see Appendix A). Museum-based participants were not paid for their participation, as is customary in museum-based labs. Online participants were compensated US$5.00 through the Prolific website. Participant demographics are in Appendix B.

2.2 Stimuli

The stimuli included 60 Hearing in Noise Test for Children (HINT-C; Nilsson et al. 1994) sentences produced by seven female talkers each representing one of seven different accent varieties: three native (Midland American English, Standard Southern British English, Scottish English), three nonnative (German-accented English, Mandarin-accented English, Japanese-accented English), and one bilingual (Hindi-English). The nonnative accents were chosen because the first languages of the speakers represent different language families and thus were likely to include distinct pronunciation features. The native and bilingual accents were selected to have talkers whose pronunciations differed from Midland American English to various extents. All speakers were recorded in a sound-attenuated booth either at Ohio State University (the Scottish English speaker) or at Indiana University (all other speakers) using a Marantz digital recorder and a Shure microphone. Root Mean Squared (RMS) amplitude was equalized across the stimuli. The Midland talker and all nonnative talkers were selected from the Hoosier Database of Native and Nonnative Speech for Children (Atagi and Bent 2013).[1]

2.3 Procedure

Participants were assigned to one of three accent conditions. All conditions included the Midland American English speaker. The accents for the other conditions were as follows: (1) Japanese-accented English and Standard Southern British English,[2] (2) German-accented English and Scottish English, and (3) Hindi-English bilingual and Mandarin-accented English.[3] The combinations of accents within each condition were designed so each listener would be presented with one nonnative accent and one native or bilingual accent that would further elicit a range of intelligibility scores. Within each of these accent conditions, participants were randomly placed in one of three noise conditions: quiet, +4 dB signal-to-noise ratio (SNR), or 0 dB SNR.[4] For the noise-added sentences, a randomly selected portion of an 8-talker babble file (Van Engen et al. 2014) was selected that was 1 s longer than the sentence with a 500 ms noise lead and tail. For each combination of accent and noise condition, there were 15–18 online participants and 26–32 in-person participants.

1 All stimuli are available in the OSF repository (https://osf.io/cnxgt/, accessed 22 April 2023). Additional materials (words, sentences, paragraphs) from the same talkers as well as materials from other talkers (with the same and different accents as included here) are freely available in the Hoosier Database for researchers and clinicians, which can be accessed via SpeechBox (Bradlow n.d.).

2 In-person data for this condition has been reported previously (Bent and Holt 2018).

3 In-person data for the condition with German-accented English and Scottish English as well as the condition with Mandarin-accented English and Hindi-English bilingual English for the quiet and the +4 dB SNR has been previously reported (Bent et al. 2021). The data for the 0 dB SNR for these two conditions and the online data have not been previously published.

4 Python code is available in the OSF repository for mixing sentences with noise, if researchers want to create new stimuli similar to those used here.

Before the start of the experimental trials, participants were presented with nine practice trials that included three sentences from each of the three talkers included in their assigned accents and noise condition. Then listeners were presented with 60 experimental trials including 20 sentences from each of the talkers within their assigned condition. The trials were blocked by talker, and the sentences were randomized within a block. The assignment of sentences to talker was counterbalanced across listeners. No feedback was provided to listeners. Consistent with the HINT-C scoring guidelines, participants were not penalized for the following substitutions: *the*/*a*/*an*, *is*/*was*, *are*/*were*, *has*/*had*, and *had*/*have*.

In-person participants were tested in a laboratory space that is not sound-treated but is separate from museum noise and generally quiet. The experimenter only carried out the procedure when the ambient noise levels were low. The stimuli were presented at a comfortable listening level binaurally over Audiotechnica headphones (model 8TH-770COM). Stimulus presentation was controlled by E-Prime (version 2.0; 2007) on a Dell Optiplex 790 desktop computer. After the presentation of each sentence, listeners repeated what they heard aloud, and the experimenter scored their response in real time by writing out each response and indicating which words were not correctly perceived (see Bent and Atagi [2017] for reliability of this scoring method).

Online participants were recruited through Prolific. If they met the inclusion and exclusion criteria, the study would appear as one for which they were eligible and they could opt to participate. All included participants passed a headphone screening (Woods et al. 2017), involving six trials. Each trial contains a series of three pure tones, one of which is 180 degrees out of phase across the stereo channels, resulting in phase cancellation. The listener is instructed to select the quietest tone. The task should be relatively easy if the participant is wearing headphones, but difficult if listening over a loudspeaker. Listeners were provided with three opportunities to complete the screening and could not continue if they did not successfully complete it. Online participants reported that the noise level in their environment was 1.99 (range = 1–8) on a scale of 1–10 (1 = very quiet; 10 = very loud). Participants used their own computers and headphones. The experiment was programmed in PsychoPy (version 2020.1) and run through Pavlovia, the online platform for PsychoPy (Peirce et al. 2019). Participants typed in what they heard. A custom Python script was run for automated scoring of participants' responses.

3 Results

The primary statistical analysis was designed to determine how word recognition accuracy differed across the two settings, for a range of accents and a range of noise conditions. A logistic mixed-effects regression model was constructed to predict word recognition accuracy with accent, noise condition, setting (in-person, online), and their interactions as fixed effects using the lme4 package for R version 4.1.1 (Bates et al. 2015; R Core Team 2021).[5] The fixed effects were treatment-coded with the following reference levels: Midland for speaker accent; quiet for noise condition; and in-person for setting. The maximal random effect structure that achieved convergence was used and included a random intercept for items (i.e., stimulus words). Statistical significance for interactions between the predictor variables was assessed using log-likelihood comparison of nested models, and significance of individual levels of the predictor variables were assessed with pairwise comparisons using the emmeans package in R.

Average accuracy and individual listener scores across noise conditions, speaker accents, and settings are displayed in Figure 1. Slope estimates, z ratios, and p values for all simple effects and interactions from the mixed-effects model are in Appendix C. The three-way interaction between noise condition, setting, and speaker accent was assessed for significance using log-likelihood comparisons of nested models, with the only difference between the models being the inclusion or exclusion of the three-way interaction. This interaction was significant, χ^2 (12) = 140.46, $p < 0.001$.

To determine how the effect of setting varied across accents and noise conditions, pairwise comparisons between online and in-person settings were carried out for each accent in each of the three noise conditions. The setting slope estimates for each accent at each noise condition are listed in Table 1. Word recognition accuracy by

5 Statistical code for the model and associated data file are available in the OSF repository.

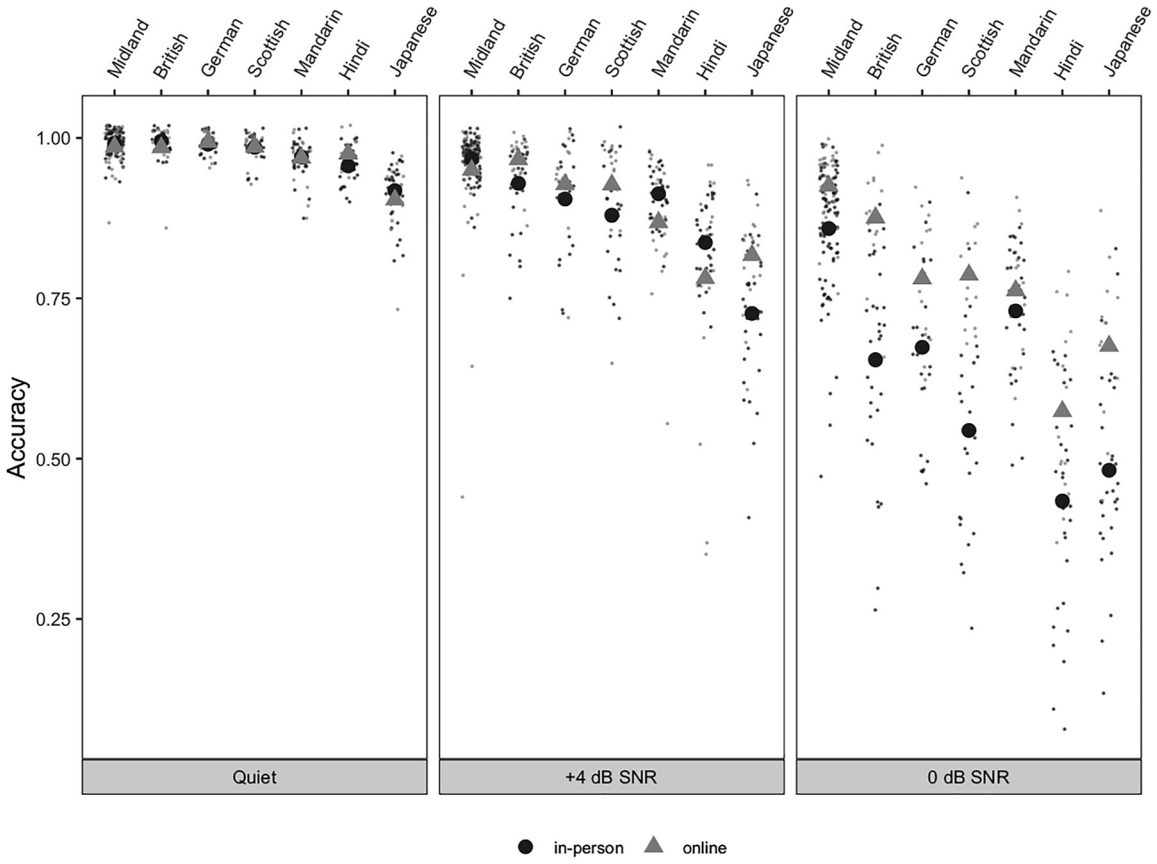

Figure 1: Listener accuracy in proportion correct across noise condition and speaker accent as a function of setting. Large black circles and gray triangles represent mean performance; small black dots and gray dots represent individual performance, for in-person testing and online testing, respectively. Individual panels represent the three noise conditions. Speaker accent is indicated on the *x*-axis.

online participants was significantly higher than those tested in person for all accents in 0 dB SNR. For +4 dB SNR, there were some exceptions to the online advantage; specifically, in-person performance was significantly higher than online for Midland, Hindi, and Mandarin accents. Similarly, in-person participants outperformed online participants in the quiet condition for Midland, Standard Southern British, and Japanese accents, while online participants demonstrated significantly higher accuracy for the Hindi accent. Differences between settings in the quiet condition should be interpreted with caution, however, given that performance by participants in both settings is near ceiling. Despite higher accuracy for most accents for the online participants, the overall performance trend across accents in the online setting generally follows that for the in-person setting (Figure 1).

Table 1: Slope estimates for pairwise comparisons of online versus in-person testing for each accent in quiet, +4, and 0 dB SNR. Negative estimates reflect better performance online than in person.

Dialect	Quiet			+4 dB SNR			0 dB SNR		
	Estimate	*z* ratio	*p* value	Estimate	*z* ratio	*p* value	Estimate	*z* ratio	*p* value
Midland	0.51	3.40	<0.001	0.45	5.25	<0.0001	−0.76	−11.70	<0.0001
Standard Southern British	1.12	3.74	<0.001	−0.79	−5.46	<0.0001	−1.49	−16.71	<0.0001
German	−0.42	−1.21	0.23	−0.33	−2.76	0.006	−0.63	−8.05	<0.0001
Scottish	−0.01	−0.03	0.98	−0.61	−5.31	<0.0001	−1.33	−17.01	<0.0001
Mandarin	0.05	0.29	0.77	0.52	5.35	<0.0001	−0.20	−2.59	0.001
Hindi	−0.60	−3.36	0.001	0.40	5.14	<0.0001	−0.67	−9.78	<0.0001
Japanese	0.21	2.01	0.04	−0.61	−8.16	<0.0001	−0.95	−13.67	<0.0001

Beyond accuracy, one difference that was noted between the two settings was that museum-based participants were significantly more likely to not respond on a trial compared to online participants, Welch two-sample *t*-test: $t(129.53) = 8.057$, $p < 0.0001$, $d = -1.23$ (in-person, $M = 5.68$; online, $M = 0.95$).

4 Discussion

The central issue addressed in this study was whether intelligibility patterns across noise conditions and accents are similar for participants tested in person compared to an online setting. For the quiet condition, there were significant differences for four of the seven accents with in-person participants outperforming online participants on three (Midland, Standard Southern British, and Japanese-accented) and online participants outperforming in-person participants on one (Hindi-English bilingual). However, performance on all accents except the Japanese-accented talker was close to ceiling in both settings. Therefore, when participants were presented with stimuli that were not mixed with noise, even if the speaker had an unfamiliar accent, intelligibility differences between the two settings were small. The finding that the online participants showed high accuracy across nearly all accents in the quiet condition suggests that environmental or technology-related factors for online participants such as ambient noise levels, interruptions, distractions, headphones, or operating systems do not appear to strongly impact performance for a relatively easy task.

In contrast to the quiet condition, there were larger accuracy differences between the two settings for the noise-added conditions. In the easier signal-to-noise ratio (i.e., +4 dB SNR), intelligibility scores averaged across accents for the two settings were similar (online: 90.3 %, in-person: 89.8 %) and there was not a consistent advantage for either setting. Museum-tested participants outperformed online participants for three of the accents (Midland, Mandarin, and Hindi) and online participants outperformed museum participants for the other four accents (Standard Southern British, German, Scottish, and Japanese). For the most difficult noise condition (0 dB SNR), the online participants showed a consistent advantage for all accents, although the gap in performance differed depending on accent with a relatively small difference between the two settings for some accents (e.g., Mandarin = 76 vs. 73 %) and large differences for others (e.g., Standard Southern British = 88 vs. 65 %; Scottish = 79 vs. 54 %). For listeners from both settings, intelligibility was highest for the Midland talker and lowest for the Japanese-accented and Hindi-English talkers with the Hindi-English talker showing the lowest performance for both groups. Conclusions regarding accent intelligibility more broadly cannot be made from this study because there was only one talker per accent. Talkers from the accents represented here but with different levels of proficiency (for the L2 talkers), residential histories, or language learning profiles, among other factors, are likely to differ in intelligibility.

It is not clear what is driving the online advantage for the most difficult noise condition, particularly because previous studies generally have shown a disadvantage for listeners tested online compared to in-lab testing (Cooke et al. 2011; Mayo et al. 2012; Slote and Strand 2016; Wolters et al. 2010) or similar performance across settings (Cooke and Garcia Lecumberri 2021). There are no obvious demographic or experience-based explanations for the performance differences. Both groups reported similar levels of exposure to the accents in their conditions, with the only large difference in exposure ratings appearing for the Midland accent. It is not surprising that the participants tested in the museum reported greater exposure to the Midland dialect, as the museum is in the Midland dialect region. In contrast, the online participants were more widely distributed across dialect regions because online participants could participate from any location within the United States. However, exposure to the Midland dialect did not appear to have a substantial impact on performance across the two groups for the Midland speaker. The in-person participants showed a slight advantage in the +4 dB SNR condition for the Midland dialect, but the online participants showed higher intelligibility for this speaker in the 0 dB SNR condition. The finding that non-Midland listeners still show high word identification accuracy for this talker may have arisen because this talker's speech aligned with General American English (Labov 1998), which tends to be highly intelligible to American listeners regardless of their residential history within the USA (Clopper and Bradlow 2008), likely due to its ubiquity in mainstream media.

Although we did not observe substantial differences in the dialect exposure ratings, it is possible that there was variability in accent exposure that was not captured in the self-reported ratings. For example, participants tested online could live in more diverse communities than those tested at the museum and therefore they may be exposed to a wider range of accent variability in their day-to-day lives. Relatedly, the participants who were from dialect regions outside the Midland region, particularly those that are more marked (e.g., the South, New England), may have consistent exposure to two native varieties that could engender broad advantages for speech processing in noise (Clopper 2014). Specifically, only 10 % of our online participants reported speaking with a Midland accent compared with 69 % of participants tested in person. Some non-Midland participants may therefore have experience with both their own dialect and General American English, which gives them a degree of perceptual flexibility that is not present for listeners who primarily communicate with other Midland speakers, even if they do not have explicit exposure to the test accents. Future work should include samples that are more closely matched on region of origin to explicitly examine whether the listener's dialect influenced the advantages observed for the online sample here. Additionally, incorporating questions about linguistic diversity in the listeners' communities could provide insight regarding how everyday exposure impacts perception of unfamiliar accents.

If researchers want participant samples that are more diverse and representative of the wider population, it is much easier to achieve this goal with online testing. In addition to greater dialect variability for online participants, the online sample was also more racially and ethnically diverse compared to the museum sample. There was a greater proportion of participants who identified as Asian or Hispanic/Latinx in the online sample compared to the museum sample. However, participants identifying as Black/African American or Hispanic/Latinx were still underrepresented relative to the US population (cf. Levay et al. 2016). Although neither of our samples fully captured the racial and ethnic diversity of the broader population, both data collection sites have an advantage over the typical populations assessed in university-based laboratories. Future work should continue to strive for participant samples that are more representative of the population.

Although online testing has the advantage of participant samples that tend to be more representative of the general population in terms of race, ethnicity, educational attainment, geographic location, and other demographic characteristics, researchers have raised concerns about online testing that could outweigh these advantages. However, some of these concerns, particularly technological ones, may no longer be a substantial hindrance. Most studies reporting intelligibility disadvantages for online participants collected the data at least 5–10 years ago (Cooke et al. 2011; Mayo et al. 2012; Slote and Strand 2016; Wolters et al. 2010). Software and hardware advances as well as the ubiquity of higher-quality headphones over the past decade may have decreased the discrepancy between laboratory-based versus personal equipment. Furthermore, the ambient noise levels for online testing did not appear to impact performance since generally participants tested at home outperformed those tested in the museum (Cooke and Garcia Lecumberri 2021; Merchant et al. 2021). Finally, our instructions were easy to understand. Therefore, task understanding was unlikely to have impacted performance across settings, although this concern should be considered for more complex experimental tasks.

One factor that could have influenced performance was the presence of the experimenter for the in-person setting compared to the online setting. Specifically, the in-person participants repeated what they heard aloud to the experimenter, who then scored their response. The online participants typed in their response. Museum participants verbally indicated their responses because adults' performance was compared to children's in another study. Since children could not type their responses, we opted to keep the response modality the same for the children and adults tested in the museum laboratory. However, online participants could not verbally report their responses to an experimenter and therefore we had them type their responses. This difference in response method, and specifically the presence of an experimenter, may have resulted in changes in listener strategy, especially when the task was difficult, as in the 0 dB SNR condition. There are a range of research participation effects (e.g., McCambridge et al. 2014) that can change or

bias behavior in experiments including demand characteristics (i.e., participants may try to fulfill experimenter implicit preferences rather than following explicit instructions; McCambridge et al. 2012). Demand characteristics could have shifted response strategies in ways that disadvantaged the in-person group even though both participant groups were encouraged to provide their best guess through written and/or verbal instructions. That is, in-person participants may have felt pressure to provide correct responses so that they would be perceived as good subjects rather than truly providing their best guess, even if their response seemed highly implausible. This shift in response strategy may have been especially prominent in the most difficult noise condition in which listeners were likely to experience much more uncertainty about their response. In contrast, participants tested online could type any response without concern about whether their response diverged substantially from the target sentence. Providing responses to more trials may have provided the online listeners an advantage over the in-person listeners. To determine whether the experimenter presence underlies the performance differences across settings, an additional set of museum-based participants could be tested who type in their responses. This change in response modality may not completely remove the impact of demand characteristics since the participant would still be interacting with the experimenter, but there would be a greater degree of separation between the response provision and the experimenter's observation of the response. Prior intelligibility studies comparing in-person and online modalities have required all participants to type their responses (Cooke and Garcia Lecumberri 2021; Cooke et al. 2011; Mayo et al. 2012; Slote and Strand 2016; Wolters et al. 2010) and have not observed better performance for online participants than in-person ones, suggesting that response modality may have contributed to the in-person performance decrements.

The only other study to directly compare in-person to online testing for nonnative-accented speech in noise (Vaughn 2019) also found an advantage for online participants. As in Vaughn (2019), our online participants were also significantly older than our in-person participants, but the difference was not large compared to the age difference in Vaughn (2019); our participants were within the same age range (i.e., 18–35) with very similar average ages (24 vs. 26 years). It should also be noted that in Vaughn (2019), nearly two-thirds of the online participants were excluded from the final data set for not attending to the instructions, failing attention checks, or not fitting inclusion criteria.

One difference for this study in contrast to most of the online versus in-person study comparisons is that our in-person group was tested in a museum-based setting rather than a traditional university laboratory. Other work that has directly compared data collected in museums with university laboratories found some performance differences between the two settings. Specifically, lower performance on a word recognition in noise task (Jones and Clopper 2019) and slower response times for a prosodic contour processing task (Ito et al. 2017) have been observed for museum-tested compared to university laboratory-based participants. Thus, performance decrements for our in-person sample could have been linked to factors related to the museum setting itself including more noise and distractions than a traditional university laboratory (Ito et al. 2017). When comparing university- and museum-based laboratories, it is straightforward to determine which setting has more noise and distractions; for the museum versus online settings used here, we cannot determine which one had more noise or distractions since we do not have objective measurements of these variables.

The motivation for participating in the experiment could also differ across modalities. Specifically, many laboratory-based studies use college students who are completing the study as a course requirement or are getting paid to participate. In contrast, participants in the museum did not receive payment and still participated, largely because visitors to a science museum have an interest in science; participation provides them with a real-life experience as a citizen scientist and an opportunity to contribute to scientific knowledge. Finally, although participants recruited from platforms such as Prolific or MTurk are likely completing the experiments for monetary compensation, there may be additional motivation to complete the task to the best of their ability since high approval ratings may lead to more and higher-paid opportunities on the platform. Future work could explicitly manipulate whether listeners are compensated for their participation to determine how this factor impacts performance.

In conclusion, these results provide support for using online data collection methodologies for word recognition tasks. Although researchers should be aware that absolute intelligibility levels may differ across testing settings, particularly for difficult tasks, relative difficulty across talkers appears stable.

Appendix A: Exclusion criteria

Most exclusion criteria for participants who completed the entire protocol were applicable for both in-person and online participants, including not meeting the language background criteria (museum: n = 8; online: n = 2), frequent exposure to the accents other than Midland American English included in the experiment (i.e., a rating of 5 on a scale of 1–5 where 1 = no exposure and 5 = frequent daily exposure; museum: n = 14; online: n = 3), self-reported speech/hearing problems (museum: n = 2) and technical problems with the software (museum: n = 4; online: n = 4). Other criteria only applied to specific contexts. Specifically, online participants were asked if they turned off all devices prior to the start of the experiment; they were excluded if they reported that they did not turn off other devices (n = 2). The online participants identified as not complying with the task did not provide responses for more than 10 % of trials for the Midland American English speaker (n = 6). Not fitting the age range criteria (i.e., not being between 18 and 35 years old) only impacted in-person participants (n = 2) because the recruitment platform used for online recruitment automatically screens out participants who are outside of the specified age range. Online participants who did not pass the headphone screening could not continue to the experimental task (n = 28), but this criterion did not impact in-person participants who were provided with headphones. An additional 21 online participants did not complete the intelligibility task due to either experimenter error (n = 2) or withdrawing following completion of the demographic survey (n = 19). Thus, combined with the headphone screening failures, a total of 49 online participants began the study but did not complete the full protocol (accounting for 23 % of online participants who began the study).

Appendix B: Participant demographics

Participant demographics for participants tested in the museum and online are provided here. Tests of group differences on age, race, and ethnicity variables between the in-person and online participants are also shown. Dialect is based on participant self-report of their regional dialect, which was indicated by the participant selecting a dialect region on a map which best matched their regional accent.[6]

		Museum	Online	Group differences
Total (in data set)		264	144	
Age (years)		M = 23.9; range = 18–35	M = 25.8; range = 18–35	t (264.74) = 4.41, p < 0.001, d = 0.474
Race	White American	86 % (n = 227)	77 % (n = 111)	χ^2 (6) = 19.34, p = 0.004
	Black or African American	7 % (n = 19)	8 % (n = 12)	
	Asian American	1 % (n = 3)	8 % (n = 12)	
	Bi- or multiracial	3 % (n = 9)	4 % (n = 6)	
	American Indian or Alaska Native	0 % (n = 0)	0 % (n = 0)	
	Native Hawaiian or other Pacific Islander	0 % (n = 0)	0.7 % (n = 1)	
	Other	0 % (n = 0)	0.7 % (n = 1)	
	Prefer not to say	2 % (n = 6)	0.7 % (n = 1)	
Ethnicity	Hispanic/Latinx	3 % (n = 7)	10 % (n = 14)	χ^2 (2) = 17.25, p < 0.001
	Not Hispanic/Latinx	92 % (n = 242)	90 % (n = 130)	
	Prefer not to say	6 % (n = 15)	0 % (n = 0)	

6 Given that folk linguistic dialect regions do not correlate directly with the regions defined by sociolinguists based on authentic production (Niedzielski and Preston 2000), this self-reported dialect may reflect where participants identify as being from more than their accent. However, our goal was to capture dialect exposure and this kind of regional identity provides a good index of primary dialect exposure.

(continued)

		Museum	Online	Group differences
Dialect	Midland	69 % (*n* = 183)	10 % (*n* = 15)	χ^2 (8) = 206.23, p < 0.001
	North	14 % (*n* = 36)	14 % (*n* = 20)	
	South (including Florida)	6 % (*n* = 17)	29 % (*n* = 42)	
	West	2 % (*n* = 5)	27 % (*n* = 39)	
	Mid-Atlantic	1 % (*n* = 3)	12 % (*n* = 17)	
	New England	0 % (*n* = 0)	5 % (*n* = 7)	
	Western Pennsylvania	2 % (*n* = 6)	3 % (*n* = 4)	
	Multiple or other	4 % (*n* = 10)	0 % (*n* = 0)	
	Chose not to respond	2 % (*n* = 4)	0 % (*n* = 0)	
Gender	Woman	51 % (*n* = 134)	48 % (*n* = 69)	χ^2 (2) = 13.06, p = 0.001
	Man	49 % (*n* = 130)	47 % (*n* = 67)	
	Nonbinary	0 % (*n* = 0)	5 % (*n* = 7)	
	Other	0 % (*n* = 0)	0.7 % (*n* = 1)	
Dialect exposure	Midland American English	4.8	2.7	t (406) = −19.77, p < 0.001, d = 2.05
	Standard Southern British English	1.9	2.0	t (143) = −0.052, p = 0.96, d = 0.01
	Scottish English	1.2	1.6	t (121) = 3.70, p < 0.001, d = 0.69
	German-accented English	1.3	1.8	t (122) = 3.59, p < 0.001, d = 0.668
	Indian English or Hindi-accented English	1.5	1.8	t (137) = 2.25, p = 0.026, d = 0.40
	Japanese-accented English	1.4	1.6	t (143) = 1.57, p = 0.12, d = 0.276
	Mandarin-accented English	1.5	1.5	t (137) = −0.08, p = 0.93, d = 0.015

Appendix C

Summary of the logistic mixed-effects model predicting word accuracy from listener setting, talker dialect, noise condition, and their interactions.

	Estimate	*SE*	*z* ratio	*p* value
Intercept	5.18	0.12	44.32	<0.0001
Variety (reference = Midland)				
Standard Southern British	0.53	0.25	2.12	0.03
German	−0.18	0.21	−0.86	0.39
Scottish	−0.57	0.18	−3.12	0.002
Mandarin	−1.33	0.15	−9.04	<0.0001
Hindi	−0.174	0.14	−12.80	<0.0001
Japanese	−2.44	0.12	−19.97	<0.0001
Condition (reference = quiet)				
0 dB SNR	−3.10	0.11	−28.38	<0.0001
+4 dB SNR	−1.42	0.12	−11.98	<0.0001
Setting (reference = in-person)				
Online	−0.51	0.15	−3.40	<0.001
Variety × condition				
Standard Southern British × 0 dB SNR	−1.87	0.26	−7.28	<0.0001
German × 0 dB SNR	−1.07	0.22	−4.91	<0.0001
Scottish × 0 dB SNR	−1.29	0.19	−6.76	<0.0001
Mandarin × 0 dB SNR	0.41	0.16	2.63	0.01
Hindi × 0 dB SNR	−0.66	0.15	−4.55	<0.0001
Japanese × 0 dB SNR	0.28	0.13	2.10	0.04
Standard Southern British × +4 dB SNR	−1.39	0.27	−5.20	<0.0001
German × +4 dB SNR	−1.03	0.23	−4.54	<0.0001

(continued)

	Estimate	*SE*	*z* ratio	*p* value
Scottish × +4 dB SNR	−0.93	0.20	−4.67	<0.0001
Mandarin × +4 dB SNR	0.23	0.17	1.34	0.18
Hindi × +4 dB SNR	−0.14	0.15	−0.91	0.36
Japanese × +4 dB SNR	−0.18	0.14	−1.31	0.19
Variety × setting				
Standard Southern British × online	−0.60	0.33	−1.85	0.06
German × online	0.93	0.36	2.56	0.01
Scottish × online	0.52	0.29	1.80	0.07
Mandarin × online	0.46	0.23	2.03	0.04
Hindi × online	1.11	0.23	4.84	<0.0001
Japanese × online	0.30	0.18	1.63	0.10
Condition × setting				
Online × 0 dB SNR	1.28	0.16	7.81	<0.0001
Online × +4 dB SNR	0.06	0.17	0.37	0.71
Variety × setting × condition				
Standard Southern British × 0 dB SNR × online	1.33	0.34	3.91	<0.0001
German × 0 dB SNR × online	−1.06	0.38	−2.81	0.005
Scottish × 0 dB SNR × online	0.04	0.31	0.15	0.88
Mandarin × 0 dB SNR × online	−1.03	0.25	−4.14	<0.0001
Hindi × 0 dB SNR × online	−1.21	0.25	−4.90	<0.0001
Japanese × 0 dB SNR × online	−0.11	0.21	−0.54	0.59
Standard Southern British × +4 dB SNR × online	1.84	0.36	5.09	<0.0001
German × +4 dB SNR × online	−0.15	0.39	−0.39	0.69
Scottish × +4 dB SNR × online	0.54	0.32	1.69	0.09
Mandarin × +4 dB SNR × online	−0.53	0.26	−2.06	0.04
Hindi × +4 dB SNR × online	−1.07	0.25	−4.20	<0.0001
Japanese × +4 dB SNR × online	0.76	0.21	3.57	<0.001

References

Adank, Patti, Bronwen G. Evans, Jane Stuart-Smith & Sophie K. Scott. 2009. Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance* 35(2). 520–529.

Atagi, Eriko & Tessa Bent. 2013. Auditory free classification of nonnative speech. *Journal of Phonetics* 41(6). 509–519.

Balota, David A., Maura Pilotti & Michael J. Cortese. 2001. Subjective frequency estimates for 2,938 monosyllabic words. *Memory & Cognition* 29(4). 639–647.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48.

Bent, Tessa & Eriko Atagi. 2017. Perception of nonnative-accented sentences by 5- to 8-year-olds and adults: The role of phonological processing skills. *Language and Speech* 60(1). 110–122.

Bent, Tessa, Melissa Baese-Berk, Stephanie A. Borrie & Megan McKee. 2016. Individual differences in the perception of regional, nonnative, and disordered speech varieties. *Journal of the Acoustical Society of America* 140(5). 3775–3786.

Bent, Tessa & Rachael F. Holt. 2018. Shhh … I need quiet! Children's understanding of American, British, and Japanese-accented English speakers. *Language and Speech* 61(4). 657–673.

Bent, Tessa, Rachael F. Holt, Kristin J. Van Engen, Izabela A. Jamsek, Lian J. Arzbecker, Laura Liang & Emma Brown. 2021. How pronunciation distance impacts word recognition in children and adults. *Journal of the Acoustical Society of America* 150(6). 4103–4117.

Borrie, Stephanie A., Melissa Baese-Berk, Kristin Van Engen & Tessa Bent. 2017. A relationship between processing speech in noise and dysarthric speech. *Journal of the Acoustical Society of America* 141(6). 4660–4667.

Bradlow, Ann R. n.d. SpeechBox. https://speechbox.linguistics.northwestern.edu/ (accessed 22 April 2023).

Buhrmester, Michael, Sanaz Talaifar & Samuel D. Gosling. 2018. An evaluation of Amazon's Mechanical Turk, its rapid rise, and its effective use. *Perspectives on Psychological Science* 13(2). 149–154.

Buhrmester, Michael, Tracy Kwang & Samuel D. Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1). 3–5.

Burchill, Zachary, Linda Liu & T. Florian Jaeger. 2018. Maintaining information about speech input during accent adaptation. *PLoS One* 13(8). e0199358.

Casler, Krista, Lydia Bickel & Elizabeth Hackett. 2013. Separate but equal? A comparison of participants and data gathered via Amazon's MTurk, social media, and face-to-face behavioral testing. *Computers in Human Behavior* 29(6). 2156–2160.

Clopper, Cynthia G. 2014. Sound change in the individual: Effects of exposure on cross-dialect speech processing. *Laboratory Phonology* 5(1). 69–90.

Clopper, Cynthia G. & Ann R. Bradlow. 2008. Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech* 51(3). 175–198.

Cooke, Martin, Jon Barker, Maria Luisa Garcia Lecumberri & Krzysztof Wasilewski. 2011. Crowdsourcing for word recognition in noise. *Proceedings of the 12th annual conference of the International Speech Communication Association (InterSpeech 2011)*, 3049–3052. Florence, Italy: International Speech Communication Association.

Cooke, Martin & Maria Luisa Garcia Lecumberri. 2021. How reliable are online speech intelligibility studies with known listener cohorts? *Journal of the Acoustical Society of America* 150(2). 1390–1401.

Cooper, Angela & Ann Bradlow. 2018. Training-induced pattern-specific phonetic adjustments by first and second language listeners. *Journal of Phonetics* 68. 32–49.

Cronk, Brian C. & Jamie L. West. 2002. Personality research on the internet: A comparison of web-based and traditional instruments in take-home and in-class settings. *Behavior Research Methods, Instruments & Computers* 34(2). 177–180.

Crump, Matthew J. C., John V. McDonnell & Todd M. Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLoS One* 8(3). e57410.

Dandurand, Frédéric, Thomas R. Shultz & Kristine H. Onishi. 2008. Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods* 40(2). 428–434.

Dragojevic, Marko & Howard Giles. 2016. I don't like you because you're hard to understand: The role of processing fluency in the language attitudes process. *Human Communication Research* 42(3). 396–420.

E-Prime, version 2.0 [Computer program]. 2007. Pittsburgh, PA: Psychology Software Tools.

Floccia, Caroline, Joseph Butler, Jeremy Goslin & Lucy Ellis. 2009. Regional and foreign accent processing in English: Can listeners adapt? *Journal of Psycholinguistic Research* 38(4). 379–412.

Ito, Kiwako, Rory Turnbull & Shari R. Speer. 2017. Allophonic tunes of contrast: Lab and spontaneous speech lead to equivalent fixation responses in museum visitors. *Laboratory Phonology* 8(1). 6.

Jiao, Yishan, Amy LaCross, Visar Berisha & Julie Liss. 2019. Objective intelligibility assessment by automated segmental and suprasegmental listening error analysis. *Journal of Speech, Language, and Hearing Research* 62(9). 3359–3366.

Jones, Zack & Cynthia G. Clopper. 2019. Influences of listener demographics on the processing of phonetic variation. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th international congress of phonetic sciences, Melbourne, Australia 2019*, 3235–3239. Canberra: Australasian Speech Science and Technology Association.

Kleinschmidt, Dave F. & T. Florian Jaeger. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review* 122(2). 148–203.

Labov, William. 1998. The three dialects of English. In Michael D. Linn (ed.), *Handbook of dialects and language variation*, 39–81. San Diego: Academic Press.

Lansford, Kaitlin L., Stephanie A. Borrie & Lukas Bystricky. 2016. Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria. *American Journal of Speech-Language Pathology* 25(2). 233–239.

Levay, Kevin E., Jeremy Freese & James N. Druckman. 2016. The demographic and political composition of Mechanical Turk samples. *SAGE Open* 6(1). 1–17.

Liu, Linda & T. Florian Jaeger. 2018. Inferring causes during speech perception. *Cognition* 174. 55–70.

Mayo, Catherine, Vincent Aubanel & Martin Cooke. 2012. Effect of prosodic changes on speech intelligibility. *Proceedings of the 13th annual conference of the International Speech Communication Association (Interspeech 2012)*, 1708–1711. Portland, OR, USA: International Speech Communication Association.

McCambridge, Jim, Kypros Kypri & Diana Elbourne. 2014. Research participation effects: A skeleton in the methodological cupboard. *Journal of Clinical Epidemiology* 67(8). 845–849.

McCambridge, Jim, Marijn de Bruin & John Witton. 2012. The effects of demand characteristics on research participant behaviours in non-laboratory settings: A systematic review. *PLoS One* 7(6). e39116.

Melguy, Yevgeniy Vasilyevich & Keith Johnson. 2021. General adaptation to accented English: Speech intelligibility unaffected by perceived source of non-native accent. *Journal of the Acoustical Society of America* 149(4). 2602–2614.

Merchant, Gabrielle R., Claire Dorey, Heather L. Porter, Emily Buss & Lori J. Leibold. 2021. Feasibility of remote assessment of the binaural intelligibility level difference in school-age children. *JASA Express Letters* 1(1). 014405.

Niedzielski, Nancy A. & Dennis R. Preston. 2000. *Folk linguistics*. Berlin: Mouton de Gruyter.

Nilsson, Michael, Sigfrid D. Soli & Jean A. Sullivan. 1994. Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America* 95(2). 1085–1099.

Pare, Dwayne E. & George S. Cree. 2009. Web-based image norming: How do object familiarity and visual complexity ratings compare when collected in-lab versus online? *Behavior Research Methods* 41(3). 699–704.

Peer, Eyal, Laura Brandimarte, Sonam Samat & Alessandro Acquisti. 2017. Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70. 153–163.

Peirce, Jonathan, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Hochenberger, Hiroyuki Sogo, Erik Kastman & Jonas K. Lindelov. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods* 51(1). 195–203.

Pichora-Fuller, M. Kathleen, Sophia E. Kramer, Mark A. Eckert, Brent Edwards, Benjamin W. Y. Hornsby, Larry E. Humes, Ulrike Lemke, Thomas Lunner, Mohan Matthen, Carol L. Mackersie, Graham Naylor, Natalie A. Phillips, Michael Richter, Mary Rudner, Mitchell S. Sommers, Kelly L. Tremblay & Arthur Wingfield. 2016. Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL). *Ear and Hearing* 37. 5S–27S.

R Core Team. 2021. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.R-project.org/.

Ramsey, Sarah R., Kristen L. Thompson, Melissa McKenzie & Alan Rosenbaum. 2016. Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior* 58. 354–360.

Rezlescu, Constantin, Iulian Danaila, Alexandru Miron & Ciprian Amariei. 2020. More time for science: Using testable to create and share behavioral experiments faster, recruit better participants, and engage students in hands-on research. *Progress in Brain Research* 253. 243–262.

Rönnberg, Jerker, Thomas Lunner, Adriana Zekveld, Patrik Sörqvist, Henrik Danielsson, Björn Lyxell, Örjan Dahlström, Carine Signoret, Stefan Stenfelt, M. Kathleen Pichora-Fuller & Mary Rudner. 2013. The Ease of Language Understanding (ELU) model: Theoretical, empirical, and clinical advances. *Frontiers in Systems Neuroscience* 7. 31.

Slote, Joseph & Julia F. Strand. 2016. Conducting spoken word recognition research online: Validation and a new timing method. *Behavior Research Methods* 48(2). 553–566.

Sprouse, Jon. 2011. A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods* 43(1). 155–167.

Van Engen, Kristin J., Jasmine E. B. Phelps, Rajka Smiljanic & Bharath Chandrasekaran. 2014. Enhancing speech intelligibility: Interactions among context, modality, speech style, and masker. *Journal of Speech, Language and Hearing Research* 57(5). 1908–1918.

Vaughn, Charlotte R. 2019. Expectations about the source of a speaker's accent affect accent adaptation. *Journal of the Acoustical Society of America* 145(5). 3218–3232.

Wagner, Laura, Shari R. Speer, Leslie C. Moore, Elizabeth A. McCullough, Kiwako Ito, Cynthia G. Clopper & Kathryn Campbell-Kibler. 2015. Linguistics in a science museum: Integrating research, teaching, and outreach at the language sciences research lab. *Language and Linguistics Compass* 9(10). 420–431.

Wolters, Maria K., Karl B. Isaac & Steve Renals. 2010. Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proceedings of the 7th speech synthesis workshop (SSW7)*, 136–141. Kyoto, Japan; International Speech Communication Association. http://hdl.handle.net/1842/4660 (accessed 22 April 2023).

Woods, Kevin J. P., Max H. Siegel, James Traer & Josh H. McDermott. 2017. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception & Psychophysics* 79(7). 2064–2072.

Xie, Xin, Kodi Weatherholtz, Larisa Bainton, Emily Rowe, Zachary Burchill, Linda Liu & T. Florian Jaeger. 2018. Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *Journal of the Acoustical Society of America* 143(4). 2013–2031.

Yoho, Sarah E. & Stephanie A. Borrie. 2018. Combining degradations: The effect of background noise on intelligibility of disordered speech. *Journal of the Acoustical Society of America* 143(1). 281–286.

Yoho, Sarah E., Stephanie A. Borrie, Tyson S. Barrett & Dane B. Whittaker. 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.