Laboratory Phonology
Journal of the Association for Laboratory Phonology

Sanker, C. (2023). How do headphone checks impact perception data? *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 14(1), pp. 1–45. DOI: https://doi.org/10.16995/labphon.8778

Open Library of Humanities

How do headphone checks impact perception data?

Chelsea Sanker, Department of Linguistics, Stanford University, Stanford, CA, USA, sanker@stanford.edu

Headphone checks have rapidly become an assumed part of best practices in online perception studies. Do they actually improve our ability to find phonological patterns? This study attempts to replicate several perceptual effects that depend on different aspects of the acoustic signal, testing whether requiring participants to pass two commonly used headphone checks (Huggins pitch perception, Milne et al., 2021; dichotic loudness perception, Woods, Siegel, Traer, & McDermott, 2017) impacts the results or the sample population being used. Participants who pass the Huggins check exhibit a larger effect of spectral tilt on perceived vowel duration, but no other effects were strengthened by either headphone check. Headphone checks actually resulted in a weaker effect of coda voicing on perceived vowel duration. A look at who is excluded by the headphone checks shows that both of them produce exclusions that are imbalanced by participants' age, gender, education, and geographic region. These imbalanced exclusions have the potential to impact experimental results in a range of ways. Thus, while headphone checks are likely to be valuable for some types of experiments, for others they might have unintended effects, and for many studies they are unlikely to have any substantial effect on the results.

1. Introduction

With increased numbers of online speech perception studies, the fields of phonetics and phonology have converged on some aspects of methodology. One of the design components that many online perception studies include is a check for whether participants are wearing headphones (e.g., Bieber & Gordon-Salant, 2022; Geller et al., 2021; Giovannone & Theodore, 2021; McPherson, Grace, & McDermott, 2022; Merritt & Bent, 2022; Saltzman & Myers, 2021; A. C. L. Yu, 2022). However, it has not quite been established how these headphone checks might be impacting different types of experiments and what the implications are for interpreting results. Understanding headphone checks is important for interpreting and comparing data collected in different ways, as well as knowing whether they should be used as criteria in evaluating research methods.

Online studies are valuable in part because of convenience; they are useful for researchers working with restrictions on in-person work, allow rapid collection of data, and make large sample sizes more feasible than they are for in-person studies. They also provide access to a diverse population of participants (Berinsky, Huber, & Lenz, 2012; Casler, Bickel, & Hackett, 2013; Shapiro, Chandler, & Mueller, 2013). Within linguistics, accessing a more diverse population is particularly relevant in facilitating research that depends on demographic groups that cannot easily come to campus, or languages and language varieties which are not spoken where the researcher is located.

Given the importance of online research, it is important to establish well-motivated best practices and to understand the effects of methodological choices. This paper focuses on headphone checks, but also touches on additional factors that can play a role in data quality, including filters for participant characteristics, instructions on required devices, and post-task questions. First, I examine whether headphone checks (or other participant filtering methods) change whether significant effects are found and what the effect sizes are. Then I discuss potential causes and implications. If different methods produce different results, then we need to consider whether the difference is a benefit or a drawback, based on who is being excluded and why. Many methodological choices for online studies are aimed at removing noisy data from participants who could not hear the stimuli or were not paying attention. However, exclusions might have other unintended effects, such as disproportionately excluding participants from certain age groups or geographic regions.

1.1. Online versus in-person data

Online studies often have slightly lower accuracy or smaller effects of experimental conditions than in-person studies. Nevertheless, there is a strong correlation between online results and in-person results, as has been demonstrated in studies that compare the same task in each modality (e.g., Cooke & García Lecumberri, 2021; Elliott, Bell, Gorin, Robinson, & Marsh, 2022; Slote & Strand, 2016; Wolters, Isaac, & Renals, 2010; A. C. L. Yu & Lee, 2014). Some studies find no significant difference between online and in-person results (e.g., Denby, Schecter, Arn, Dimov, & Goldrick, 2018).

Many of the differences between in-person and online studies seem to be due to greater variation among online participants. On the one hand, this might be due to diversity of the sample. In-person experiments often use convenience samples that are primarily university students, and the results among university students and other population groups sometimes differ (Peterson, 2001). Online studies usually recruit from a participant population that is more diverse in age and education, as well as many other characteristics (Berinsky et al., 2012; Casler et al., 2013; Shapiro et al., 2013). While these factors will not necessarily impact behavior in every linguistic study, some of them do have the potential to influence results; for example, a range of studies are likely to be impacted by hearing loss among older participants or dialect differences associated with socio-economic status. Because of these differences in the population being sampled, effects of variables like age (e.g., Shen & Wu, 2022) might appear to be effects of the online medium if a study does not control for them.

On the other hand, some of the variation in results for online participants might be due to variability in aspects of the experimental context, such as participants' listening equipment. Online studies have much less control over the experimental setup than in-person studies do. While in-person studies are often conducted in the same room with the same equipment for all participants, online participants may vary substantially in the device they use to play the stimuli and enter responses, as well as their location (both across large scale geographic areas and small-scale differences in the type of room), background noise, surrounding activity and other distractions, and stability of their internet connection. Not all of these potential sources of variation have been investigated, so it remains unclear what the relative contribution of each factor might be.

Consistent with an effect of experimental setup rather than the participants themselves, Cooke and García Lecumberri (2021) find that online participants transcribe sentences less accurately than in-person participants even when both conditions sample from the same population. They also find that self-rated headphone quality is a predictor of accuracy; online participants who reported using more expensive headphones had higher accuracy than those who reported using cheaper headphones, and those with expensive headphones performed similarly to in-person participants. There are differences in the audio produced by different broad categories like in-the-ear versus over-the-ear headphones, as well as differences within each category (Breebaart, 2017). However, variation across listening devices is not the only reason for differences between in-person and online studies. In one particularly striking example, McAllister, Preston, Ochs, and Hitchcock (2022) mailed the same model of headphones to all online participants, but still found more variability among online participants than in-person participants. Additionally, studies that do not involve audio input also often find lower accuracy or smaller effects among online participants than among in-person participants (Dandurand, Shultz, & Onishi, 2008; Schnoebelen & Kuperman, 2010).

Participants in online studies sometimes produce poor data because of low effort or limited attention. For speech perception studies, it is particularly relevant that online participants sometimes report distractions that involve speech, such as listening to music, watching television, or talking to another person (Chandler, Mueller, & Paolacci, 2014; Clifford & Jerit, 2014; Hauser, Paolacci, & Chandler, 2019). If the poor listening environment is because of additional audio that is being presented through the same device, even high quality headphones will not provide a benefit. A headphone check will usually filter out participants with low attention, low effort, or poor listening environments, but other types of checks for accuracy or attention will also filter out such participants. Some studies which use headphone checks also use additional checks for attention or audio quality, such as excluding participants with low accuracy on clear items (Brown et al., 2018; Mills, Shorey, Theodore, & Stilp, 2022; Saltzman & Myers, 2021) or inaccurate responses for catch trials that instruct them to provide a specific response (Brekelmans, Lavan, Saito, Clayards, & Wonnacott, 2022; Nayak et al., 2022; Seow & Hauser, 2022). Including multiple types of checks makes it difficult to evaluate which one(s) were most valuable in filtering out participants with low attention or poor listening environments and not filtering out additional participants. Some studies include checks to ensure attention and ability to hear the audio stimuli without including a headphone check (Denby & Goldrick, 2021; D'Onofrio, 2018; Getz & Toscano, 2021). Another approach is to include post-task questions for participants to self-report whether they were paying attention to the task, encountered technical difficulties, and/or felt their data was usable (Beier & Ferreira, 2022; Nayak et al., 2022). All of these methods are likely to filter out unusable data.

It is possible that participants in online studies are more likely to misunderstand the task than participants in in-person studies, because the experimenter is not present to verify participants' understanding. Some authors speculate that misunderstanding of the task might contribute to low accuracy among online participants. For example, in their study testing the dichotic loudness headphone check, Woods et al. (2017) suggest that the relatively high fail rate for participants who reported wearing headphones might in part be due to some of them not understanding the task instructions. However, I am not aware of any studies directly testing whether differences in understanding the task contribute to differences between in-person and online studies, or whether such an effect would differ from effects of attention.

Another potential source of variability is inaccurate information about the participants. Sometimes participants misrepresent themselves in order to access studies (Chandler & Paolacci, 2017; Peer, David, Andrew, Zak, & Ekaterina, 2021). This misrepresentation could potentially influence results due to recording inaccurate information about characteristics that are being examined as predictors, such as gender or age. It could also obscure results due to some participants not belonging to the target population; sometimes participants will claim to be native speakers of English or other languages that studies are restricted to (Aguinis, Villamor, & Ramani, 2021), which would make their data unusable for most phonological experiments. Headphone checks

will not help with this aspect of potential variation, because they are not language specific, aside from requiring participants to be able to understand the instructions.

There are many potential factors that might contribute to more variable or less accurate results in online studies, based on the experimental setup, characteristics of the participants, and how the experiments are designed. Focusing just on headphone usage might sometimes decrease how closely researchers consider other sources of variation, even though it is often unclear what the relative contribution of different factors is. Some factors are almost never considered, despite having demonstrable effects on data collection. For example, time of day impacts many demographic characteristics of the participants who do an online task (Casey, Chandler, Levine, Proctor, & Strolovitch, 2017), but this is a factor that is rarely controlled or reported.

It is also important to remember that variation across listeners and their listening environments in online studies is not necessarily a bad thing. On the one hand, it might make small effects more difficult to find, if variation across participants is essentially adding noise to the data. On the other hand, variation across participants may capture more representative patterns of behavior than would be found in a more uniform population, and variation in the experimental environment may be useful in reducing potential effects of factors like the experimenter's dialect (cf. e.g., Hay, Drager, & Warren, 2009). Online listeners may have listening environments that are more typical of natural language use, which could result in different behavior than perception under laboratory conditions; both conditions are likely to be informative (cf. lab speech versus natural speech, e.g., Wagner, Trouvain, & Zimmerer, 2015).

1.2. Checking for headphones

Many researchers want to ensure that participants use headphones to complete perception experiments. At least half of online perception studies include a headphone check to ensure that participants are wearing headphones, either the dichotic loudness test (Woods et al., 2017) or the Huggins pitch perception test (Milne et al., 2021). Following are some examples (not an exhaustive list) of online phonetics and phonology experiments divided based on whether they used headphone checks and whether they instructed listeners to use headphones. See the appendix for a tabular survey of this information.

The dichotic loudness test presented by Woods et al. (2017) is used in many experiments (e.g., Krumbiegel, Ufer, & Blank, 2022; Lavan, Knight, Hazan, & McGettigan, 2019; McPherson et al., 2022; Mepham, Bi, & Mattys, 2022; Merritt & Bent, 2022; Mills et al., 2022; Nayak et al., 2022; Saltzman & Myers, 2021; A. C. L. Yu, 2022).

The Huggins check, only recently popularized by Milne et al., 2021, already appears in many experiments (e.g., Beier & Ferreira, 2022; Brekelmans et al., 2022; Ringer, Schröger, & Grimm, 2022; Tamati, Sevich, Clausing, & Moberly, 2022; Wu & Holt, 2022).

Studies that do not include a headphone check often instruct listeners that they should use headphones; some will additionally include a question at the end of the experiment asking participants what device they used to for listening to the stimuli (Davidson, 2020; Getz & Toscano, 2021; Reinisch & Bosker, 2022), though others do not (Manker, 2020; McHaney, Tessmer, Roark, & Chandrasekaran, 2021).

However, not all studies require participants to wear headphones; some do not include a headphone check or instructions to wear headphones (e.g., Denby & Goldrick, 2021; D'Onofrio, 2018; Kato & Baese-Berk, 2022; Vujović, Ramscar, & Wonnacott, 2021; Williams, Panayotov, & Kempe, 2021). It is important to note that the lack of consistent headphone use in these studies does not seem to be an issue in finding meaningful results.

The use of headphone checks is aimed at concerns about controlling the experimental setup (Milne et al., 2021; Woods et al., 2017). It is often assumed that headphones will produce a better listening setup, more comparable to an in-person study, and that this will result in stronger or more reliable results (e.g., Brown et al., 2018; Geller et al., 2021; Seow & Hauser, 2022). However, it is often not made explicit what benefit headphones are assumed to provide. Many studies just say that the check was done to ensure that participants were wearing headphones, without a specific description of why headphones were important for the task (e.g., Beier & Ferreira, 2022; Bieber & Gordon-Salant, 2022; Brekelmans et al., 2022; Giovannone & Theodore, 2021; Krumbiegel et al., 2022; Lavan et al., 2019; Luthra et al., 2021; Mepham et al., 2022; Merritt & Bent, 2022; Mills et al., 2022; Ringer et al., 2022; Saltzman & Myers, 2021; Wu & Holt, 2022; A. C. L. Yu, 2022; M. Yu, Schertz, & Johnson, 2022). Some studies describe the headphone check as ensuring better audio conditions by improving the sound quality and decreasing back ground noise (e.g., Brown et al., 2018; Geller et al., 2021; McPherson et al., 2022; Nayak et al., 2022; Seow & Hauser, 2022). Both of the studies which provide the commonly-used headphone check methods primarily describe attenuation of environmental noise as the way that headphones are likely to improve audio conditions (Milne et al., 2021; Woods et al., 2017).

Previous work has indeed shown that headphones block out some ambient noise, though the extent of this noise attenuation varies by frequency and differs across different headphones (Ang, Koh, & Lee, 2017; Liang, Zhao, French, & Zheng, 2012; Shalool, Zainal, Gan, Umat, & Mukari, 2017). It is not clear if this is a major factor in perception studies. In a word identification task with background noise played over loudspeakers, Molesworth and Burgess (2013) found that listeners had only slightly higher accuracy when the stimuli were played over headphones with the noise cancelling function turned off than when the stimuli were played over loudspeakers, though accuracy was substantially higher with noise-cancellation turned on. The potential role of noise-cancelling headphones in online research is unclear, both because this factor has not been examined but also because the degree of noise-cancellation depends on the type of noise.

It is not clear that headphones produce higher fidelity audio than other devices. For example, low frequencies are not well represented by low quality or even average headphones, though high quality headphones perform well (Breebaart, 2017; Olive, Khonsaripour, & Welti, 2018; Wycisk et al., 2022); similar differences in the relative intensity of low frequencies are found for built-in laptop speakers as compared to loudspeakers (Wycisk et al., 2022). Whether due to differences in frequency response or other factors, the quality of headphones also impacts perceptual clarity (Cooke & García Lecumberri, 2021).

The headphone checks provided by Milne et al. (2021) and Woods et al. (2017) are both based on testing whether the audio is presented in stereo at close enough distance that phase-shifting effects are not overly attenuated. These studies demonstrate that their methods are indeed predictive of headphone use; the relationship is not absolute, but the majority of participants who are wearing headphones pass these checks and the majority of participants who are not wearing headphones fail. It is important to note that the goal of these studies is just to demonstrate that these methods can be used to check for headphone use; they are not testing whether headphone checks improve data quality. Milne et al. (2021) note that stereo audio presentation is necessary for tasks like dichotic listening and spatial manipulations; however, it will not necessarily be valuable in other tasks.

In direct comparisons, is there evidence for headphone checks affecting the results of perception experiments? Differences between online and in-person studies can be found even when the online study included a headphone check (e.g., Elliott et al., 2022), and some studies find no significant difference between online and in-person results despite not using a headphone check (e.g., Denby et al., 2018). A few studies include a headphone check and compare results based on whether it is used to exclude participants; they report that excluding participants based on the headphone check did not change the results (Ringer et al., 2022; Shen & Wu, 2022).

Headphone checks might alter the demographics of which participants are included in experiments. Even if the checks are functioning as intended and identifying participants who are using headphones, this may have unintentional side effects. Device choice is predicted by a range of demographic factors, including age, gender, educational background, and income (Haan, Lugtig, & Toepoel, 2019; Lambert & Miller, 2015; Passell et al., 2021), so checking for headphones could systematically exclude more participants from certain groups. Headphone checks might also impact participant demographics based on additional factors that influence perception; for example, hearing loss can decrease accuracy in Huggins pitch detection (Santurette & Dau, 2007). In some studies, excluding listeners with hearing loss may be desirable, whereas in others it may be undesirable (e.g., in studies looking at age-related effects).

In addition to the ways that headphone checks might impact the data, it is important to consider the ways that they are not impacting the data. As described above, there are many

factors that might vary among online participants. Headphone checks do not control everything about the audio presentation or the environment; decreased variability in one aspect of audio presentation does not mean that the audio or the listening environment will be uniform in other respects. It is perhaps because of these other sources of variability that both headphone check methods are failed by a substantial number of participants who report using headphones (Milne et al., 2021; Woods et al., 2017).

1.3. This study

This article presents an online study aimed at testing whether excluding participants based on headphone checks improve the results in three perceptual studies testing effects that depend on different aspects of the acoustic signal: Vowel duration, F0, F1, and spectral tilt. All of the effects being tested have been observed previously in in-person studies, so there are strongly expected results that should be present if the data is reliable.

The impact of the headphone checks on several demographic variables is also examined, to evaluate whether headphone checks are substantially changing the set of participants who are included in experiments.

2. Methods

The study consists of three phonological perception tasks replicating previous in-person work, two headphone checks, and a brief questionnaire.[1]

Stimuli for the three speech perception tasks were produced by two female American English speakers (one for Tasks 1 and 2, one for Task 3), elicited individually in randomized order with PsychoPy (Peirce, 2007), recorded in a quiet room with a stand-mounted Blue Yeti microphone in the Audacity software program, and digitized at a 44.1 kHz sampling rate with 16-bit quantization. The characteristics of the stimuli for each task are described in the following subsections.

Stimuli for the two headphone checks came from recent studies testing each method: Milne et al's (2021) stimuli for the Huggins pitch perception check, and Woods et al's (2017) stimuli for the dichotic loudness perception check.

[1] Data was also collected for a task in which participants identified the position of stress in two-syllable words which form stress-based minimal pairs (e.g., permit /ˈpɻmɪt/, /pɻˈmɪt/). This task was intended as an additional test of listeners' attention based on accuracy in identifying clear items (as described below, there was also a task in which participants made decisions about consonant contrasts). However, this stress perception data was not used in analysis because accuracy was low; mean accuracy was 80.3%, and only 53/120 participants had accuracy above 85%. The low accuracy might suggest that many listeners rely almost entirely on vowel quality to determine stress, given how few English words have stress contrasts without corresponding vowel quality differences. The low accuracy may also indicate that the instructions or the stimuli were unclear.

The study was run online, with participants recruited and paid through Prolific and the experiment presented through Qualtrics. Participants were 120 monolingual native speakers of American English (40 male, 75 female, five nonbinary/no response; mean age 33.6).

Prolific allows filtering based on participants' overall approval percentage in previous studies, as participants who have done well on prior tasks are likely to also do well on new tasks (Peer, Vosgerau, & Acquisti, 2014). In order to get a range of participants with different levels of attention and effort while mostly collecting higher-quality data, thirty participants were recruited with this filter set at 90%, and ninety participants were recruited with this filter set at 95%.

On Prolific, the description of a study can indicate that participants should use a particular device (though this does not do any testing to ensure that participants are following the recommendations). In order to check the effect of these recommendations, thirty of the participants recruited at the 95% approval level saw a restriction that they should be using a computer or tablet (not a phone), and sixty of them did not see any restriction about what device they should use.

Prolific also allows filtering based on nationality and location. Participants were restricted to being from the United States and currently residing in the United States.

Participants were instructed that they would hear words and sounds and make decisions about what they heard. There was a different task in each block. Within a block, the order of presentation of items was randomized.

After completing the listening tasks, participants completed a brief questionnaire, which included questions about age (free response), gender (free response), highest level of education (middle school or less, some high school, high school, some college [undergraduate], college [undergraduate], some graduate/professional school, graduate/professional school), state or region where they spent most of their childhood (free response), state or region where they currently live (free response), and what device they used to listen to the stimuli (earbuds/in-ear headphones, full-size/over-the-ear headphones, built-in phone speaker, built-in tablet speaker, built-in computer speaker, external speakers). They were permitted to skip any of these questions that they preferred not to answer.

Each task contained a small number of trials, in order to ensure that the study was brief; the median completion time for the full study was slightly over 10 minutes. The duration of online tasks is important for two main reasons. First, long tasks can have high dropout rates; almost half of respondents report that they would quit an experiment if it took over 15 minutes (Sauter, Draschkow, & Mack, 2020). Second, attention to the task decreases with longer online studies, producing a corresponding drop in data quality (Yentes, 2015). Longer studies result in larger differences between the attention of in-person and online participants, with online participants paying less attention (Goodman, Cryder, & Cheema, 2013), even when performance may be similar for short tasks (Paolacci, Chandler, & Ipeirotis, 2010).

2.1. Task 1: F0 and onset stop laryngeal contrasts

Previous work has found that a higher F0 following an onset stop increases the perception that it is voiceless (and aspirated) rather than voiced (Haggard, Ambler, & Callow, 1970). This task aims at replicating that result, providing a test of pitch perception.

Stimuli were produced by a trained phonetician targeting a stop ambiguous between the voiced and voiceless categories; the VOT of each item was then manipulated to be as close as possible to the category boundary between the voiced and voiceless stop for each item (mean 27 ms for bilabials, 31 for alveolars). There were two F0 manipulations for each item. F0 was manipulated using the Change pitch median function in the Vocal Toolkit (Corretge, 2020) in Praat. In the low F0 condition, the mean was 216 Hz and in the high F0 condition, the mean was 264 Hz; these values were chosen based on being equally distant from the speaker's natural mean F0, while still being within her natural range. There were six of these ambiguous items (e.g., *b/pest*). This produced a total of 12 stimuli.

Mixed with these test stimuli, there were also 24 filler stimuli: 12 items with onset fricative place contrasts (e.g., *heft, theft*) and 12 items with coda stop place contrasts (e.g., *bud, bug*). These served to provide a test of listeners' accuracy in identifying unambiguous items.

This produced a total of 36 items, each of which was heard a single time by each listener. The intensity of all items was normalized so that they had the same mean intensity. A list of these words is given in the appendix.

The items were presented as a list; listeners clicked on each item to hear the stimulus. Responses were given by clicking on one of two written response options. For the F0-manipulated test items, these were the respective voiced and voiceless options (e.g., *best, pest*). For the filler items, the response options were the two contrasting words, one of which was a clear match for the stimulus. The order of the two response options was balanced across participants.

All 36 items in this task had the vowel /ɛ/ or /ʌ/, and they served as the exposure condition for Task 2, which tested listeners' category boundaries between high and mid vowels. Half of the participants heard these items with raised F1 and half of the participants heard them with lowered F1 (described in more detail below).

2.2. Task 2: Vowel height category boundaries

Previous work has demonstrated that vowel category boundaries can be shifted based on exposure to manipulated vowels; exposure to a higher F1 makes listeners expect vowels in that category to have a higher F1 (Ladefoged & Broadbent, 1957), so exposure to raised/lowered F1 in /ɛ, ʌ/ should produce a corresponding shift in subsequent perception of ambiguous items on /ɛ-ɪ/ continua. This provides a test of formant perception.

As mentioned above, all of the items in Task 1 had manipulated F1, to set up the exposure conditions of raised or lowered F1 in mid vowels. The response options for each exposure item differed in the onset or the coda but had the same vowel (e.g., *best, pest*), so it was unambiguous how each vowel should be categorized. Half of the participants heard those vowels in Task 1 with raised F1 (mean 918 Hz), and half heard them with lowered F1 (mean 673 Hz), equally distant in Bark from the speaker's naturally produced mean F1 for these vowels. Manipulations for these training stimuli and for the testing stimuli were done in Praat using the Change formants function in the Vocal Toolkit (Corretge, 2020), and were manually checked to ensure that the manipulation was successful. The only change was in F1, in order to focus on a single characteristic and avoid potential confounds. Stimuli in both conditions were made from the same recordings; thus, the stimuli in the two conditions differed only in F1.

The testing stimuli were continua of ambiguous vowels based on three pairs of monosyllabic English words (*itch-etch, pit-pet, tick-tech*), differing only in the vowel, /ɪ-ɛ/. All words had four F1 manipulations: F1 matching the speaker's mean F1 for each of the two vowel qualities, and two intermediate values, equally spaced in the Bark scale. The intensity of the resulting items was then normalized so that they all had the same mean intensity. Both vowels for each contrast were used to create stimuli, producing two continua (eight items): For example, the *pit, pet* items included four manipulations made from naturally produced *pit* and four manipulations made from naturally produced *pet*, with the same F1 values in both continua. No other characteristics were altered, so the two continua for each pair often differed in F2, duration, and other characteristics. Identifications of items in these vowel continua served to test the effect of the manipulations that listeners were exposed to in the training block.

The items were presented as a list; listeners clicked on each item to hear the stimulus. Responses were given by clicking on one of the written words, which differed only in the vowel (e.g., *pit, pet*). There were three pairs of words and eight items for each pair, for a total of 24 items, each of which was heard a single time by each listener. The order of the two response options was balanced across participants.

2.3. Task 3: Spectral tilt and perceived vowel duration

Sanker (2020) demonstrated that lower spectral tilt increases perceived vowel duration; this task uses a subset of the same stimuli used in that study. While this effect is not as well-established as the previous two, it is included in order to examine an expected effect that is based on spectral tilt and thus might be particularly sensitive to device, as frequency response differs across devices. This task also tests perception of vowel duration and coda voicing.

Stimuli were based on recordings of a female American English speaker producing the words *cub* and *cup*: /kʌ/ followed by a voiced or voiceless labial stop. The words were cut at the end of

the vowel, and each beginning (originally produced with /p/ versus /b/) was spliced with two different endings: /p/ and /b/. The vowel duration was manipulated to produce four duration steps, from 146 ms to 218 ms.

Each vowel was manipulated to have two spectral tilts, using the Praat Filter(formula) function to recalculate intensities relative to frequency: High (10.5 dB) and low (3.2 dB). The filter formula used to create high spectral tilt items was: *self* / $(1 + x/100)^{0.8}$, and the filter formula used to create low spectral tilt items was: *self* * $(1 + x/100)^{-0.8}$. The intensity of the resulting items was then normalized so that they all had the same mean intensity. Note that the manipulation was relative to the original spectral tilt, so it preserved voicing-conditioned spectral tilt differences.

This produced a total of 32 items, each of which was heard a single time by each listener.

The items were presented as a list; listeners clicked on each item to hear the stimulus. Responses were given by clicking on "long" or "short" as the answer to the question "Was the vowel long or short?" The order of the two response options was balanced across participants.

2.4. Headphone checks

Two headphone checks were completed after the phonological tasks. They were not described as headphone checks; the instructions merely asked participants to evaluate the given characteristic.

One headphone check was Huggins pitch perception, based on broadband noise that is phase-shifted over a narrow band by 180 degrees between stereo channels (Milne et al., 2021); stimuli came from the supplementary materials provided with that paper. The phase-shift produces an apparent pitch at the frequency of the band with the phase-shift when the channels are independent, as in headphones; each stimulus contained three sections of noise, one of which contained this phase-shift and thus should sound like it contains a tone. Listeners were asked which of the three sections of noise contained a tone. Responses were given by selecting a button with the corresponding number (1, 2, 3). The numbers were always listed in ascending order.

The other headphone check was based on loudness perception (Woods et al., 2017); stimuli came from the supplementary materials provided with that paper. In one of the items, the sound was presented phase-shifted by 180 degrees between the stereo channels, which should make that item sound quieter than the others when presented over headphones. When presented over speakers, the phase difference is less likely to align consistently. Listeners were asked which of three sounds was the quietest. Responses were given by selecting a button with the corresponding number (1, 2, 3). The numbers were always listed in ascending order.

Audio quality and attention was also checked based on accuracy of identifications of the unambiguous items in Task 1 (e.g., *bud, bug; theft, heft*); the threshold for accuracy was 85%.

3. Results

3.1. Exclusions by device and Prolific parameters

First, recall that there were different conditions in recruitment: Thirty participants were recruited with a 90% threshold for their overall approval rate and no suggestions about what device they should use. Sixty participants were recruited with a 95% threshold for their overall approval rate and no suggestions about what device they should use. Thirty participants were recruited with a 95% threshold for their overall approval rate and a suggestion that the task should be completed on a computer or tablet (and not a phone). The incomplete balancing of participants across these conditions was done in order to reduce the number of participants with very low-effort submissions, as the main goal of the experiment was to establish effects of headphone checks, not the effect of approval rate, which has been established in previous work.

The filter based on approval rating is here considered just among participants with no suggestions about what device they should use, as no device suggestions were given to participants recruited with the lower approval rating threshold. Setting higher filters improves performance for identifying clear stimuli: 4/30 (13%) of participants recruited at the 90% approval threshold had accuracy below 85% for consonant decisions, compared to 3/60 (5%) of participants at the 95% threshold. However, the filter based on approval rating did not impact exclusions based on the headphone checks. 19/30 (63%) of participants at the 90% approval threshold failed the Huggins check, compared to 39/60 (65%) of participants at the 95% threshold. 17/30 (57%) of participants at the 90% approval threshold failed the dichotic loudness check, compared to 34/60 (57%) of participants at the 95% threshold. This seems to suggest that accuracy for clear stimuli is capturing something about attention or similar behaviors that impact a range of tasks, while the headphone checks are not related to the typical quality of a participant's engagement in studies.

The effect of instructing participants on what device they should use is considered here among participants at the 95% threshold. Telling participants that they should not use their phones did impact reported device usage. Among participants without any device restriction instructions, 18/60 (30%) reported using their phones to listen to stimuli, while none of the participants with the device restriction instructions reported using their phones. Of course, this might in part reflect a difference in reporting rather than actual usage, but it also impacted how many participants passed the headphone checks, which suggests that participants were actually less likely to use phones in this condition. Among participants who saw that the task should not be completed on a phone, 13/30 (43%) failed the Huggins check (versus 65% among those without this instruction). Among participants who saw that the task should not be completed on a phone, 14/30 (47%) failed the dichotic loudness check (versus 57% among those without this instruction).

The next consideration is how many participants would be excluded by each headphone check and by the accuracy-based audio/attention check. **Table 1** shows how these exclusions aligned. Most of the participants who had low accuracy in the consonant identification task also failed the headphone checks, but there were two participants with low accuracy on these items who passed one or both of the headphone checks. Performance in one headphone check was predictive of performance in the other, but the relationship was not absolute: 26 participants (22%) passed one headphone check but failed the other, which makes it important to evaluate the effects of each headphone check separately.

	Accuracy <85%		**Accuracy >85%**	
	Huggins Fail	**Huggins Pass**	**Huggins Fail**	**Huggins Pass**
Dichotic Loudness Fail	7	0	48	10
Dichotic Loudness Pass	1	1	15	38

Table 1: Comparison of exclusions across method.

Next, the exclusions will be divided by reported device usage. **Table 2** indicates the number of participants with each Huggins score and the device that the participant reported using for the task. Milne et al. (2021) recommend a cutoff of 6/6 accurate responses, but 5/6 is also considered here.

	<5	**5**	**6**
built-in computer speaker	21	1	5
built-in tablet speaker	2	0	0
built-in phone speaker	24	0	1
external speakers	3	0	0
over-the-ear headphones	7	1	21
in-the-ear headphones	12	2	18
device not reported	2	0	0

Table 2: Number of participants by Huggins score and reported device.

Even using the threshold of 5/6 correct, the Huggins headphone check excluded 71 of the 120 participants (59%); this is the threshold used for the rest of the analyses. It does generally seem to capture headphone use, though as noted above, it is not clear whether this restriction will improve the results. There is a noteworthy divide between over-the-ear headphones and in-the-ear headphones; more participants who reported using in-the-ear headphones were excluded.

Table 3 indicates the number of participants with each score on the dichotic loudness headphone check (see Woods et al., 2017) and the device that they reported using. The recommended threshold is 5/6 correct.

	<5	5	6
built-in computer speaker	21	1	5
built-in tablet speaker	2	0	0
built-in phone speaker	20	4	1
external speakers	3	0	0
over-the-ear headphones	8	2	19
in-the-ear headphones	9	0	23
device not reported	2	0	0

Table 3: Number of participants by dichotic loudness score and reported device.

The dichotic loudness headphone check excluded 65 of the 120 participants (54%). Like the Huggins check, it seems to generally capture headphone use. Unlike the Huggins check, there was no apparent difference between over-the-ear headphones and in-the-ear headphones.

We can also consider exclusions based on how listeners identify items that should be unambiguous, to exclude participants who cannot hear the stimuli or are not paying attention to the task. **Figure 1** shows the distribution of accuracy by participant for identifying the unambiguous items in Task 1; recall that these were binary response options differing only in one consonant, with naturally produced consonants (e.g., *heft, theft*). It is clear that there are some outliers that should be excluded, but the vast majority of participants had very high accuracy; they were paying attention to the task and could hear the stimuli.

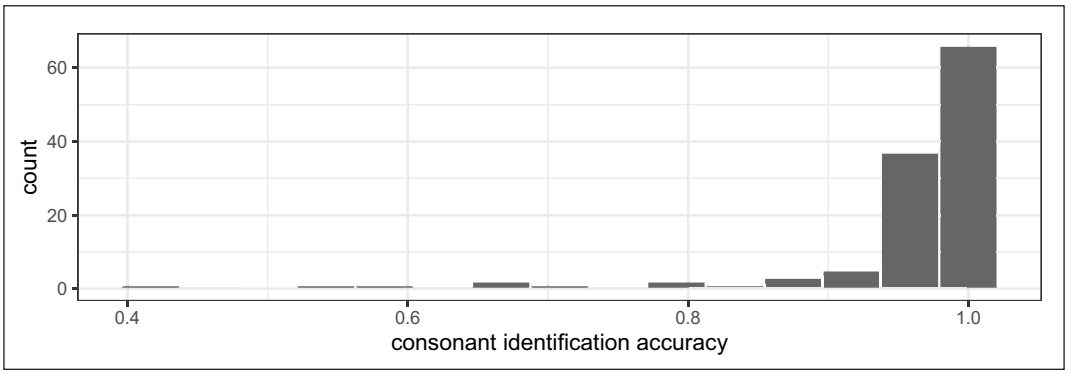

Figure 1: Proportion of accurate responses identifying the unambiguous items in Task 1, by participant.

Table 4 indicates the number of participants with each percentage of accurate identifications of the clear consonant stimuli (24 items).

	<75%	>75%, <85%	>85%
built-in computer speaker	1	1	25
built-in tablet speaker	0	0	2
built-in phone speaker	2	1	22
external speakers	1	0	2
over-the-ear headphones	0	1	28
in-the-ear headphones	2	0	30
device not reported	0	0	2

Table 4: Number of participants by consonant identification accuracy and reported device.

Using a threshold of 85% accuracy for these clear items excluded nine participants: Making the threshold higher or lower does not have a large effect on how many participants are excluded. There was no apparent relationship between accuracy and the device that the participant used.

3.2. Exclusions by demographic characteristics

Who is being excluded by the headphone checks? **Tables 5** and **6** summarize the demographic characteristics (age, gender, education, and childhood location) of participants who passed and failed each headphone check.

	mean age (SD)	**gender**	**education**	**childhood location**
Huggins Fail	34.8 (12.8)	50 f	12 high school	22 midwest
		17 m	46 undergrad	9 northeast
		4 nb/no response	13 grad/professional	23 southeast
				7 southwest
				9 west
Huggins Pass	31.8 (10.7)	25 f	13 high school	9 midwest
		23 m	32 undergrad	15 northeast
		1 nb/no response	4 grad/professional	11 southeast
				6 southwest
				7 west

Table 5: Demographics by Huggins results.

	mean age (SD)	gender	education	childhood location
Dichotic Loudness Fail	34.8 (13.6)	40 f	13 high school	16 midwest
		20 m	41 undergrad	13 northeast
		5 nb/no response	11 grad/professional	22 southeast
				7 southwest
				6 west
Dichotic Loudness Pass	32.1 (9.6)	35 f	12 high school	15 midwest
		20 m	37 undergrad	11 northeast
			6 grad/professional	12 southeast
				6 southwest
				10 west

Table 6: Demographics by Dichotic Loudness results.

In the categories for highest level of education, "some" and "completed" are merged; for example, participants who reported that they finished an undergraduate degree and participants who reported that they spent some time in an undergraduate program are reported the same way. This was done because there was little evidence for a difference between each pair of categories, and many of the people reporting "some" of a particular level of education are likely to still be in that program.

Childhood location (where the participant reported spending most of their childhood) is organized into five broad regions: Northeast (CT, DE, MA, MD, ME, NH, NJ, NY, PA, RI, VT), Southeast (AL, AR, FL, GA, KY, LA, MS, NC, SC, TN, VA, WV), Midwest (IA, IL, IN, KS, MI, MN, MO, ND, NE, OH, SD, WI), Southwest (AZ, OK, NM, TX), and West (AK, CA, CO, HI, ID, MT, NV, OR, UT, WA, WY). Two participants just put "United States" and are not reported for the childhood location column in these tables.

Both headphone checks exhibited several similar demographic patterns in who was excluded. Older participants were more likely to fail both headphone checks. Women, nonbinary people, and participants who declined to answer the question about gender were more likely to fail the headphone checks, with a particularly large difference for the Huggins check. People who had entered college or graduate/professional school were more likely to fail the headphone checks than participants whose highest level of education was high school. Participants from the west and northeast were most likely to pass the headphone checks, and participants from the southeast were least likely to pass the headphone checks.

Some of these results may reflect inherent characteristics of the participants, such as hearing loss among older participants. Other results may reflect choices about device. For example,

women were more likely to complete the task on a phone (25%) than men (10%), and people who had entered a graduate or professional school were more likely to listen to the stimuli on computer speakers (47%) than people whose highest level of education was high school (12%).

The following sections present the results for the phonological/phonetic effects being examined. Statistical results are from mixed effects models, calculated with the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015), with p-values calculated by the lmerTest package (Kuznetsova, Bruun Brockhoff, & Haubo Bojesen Christensen, 2015).

3.3. Task 1: F0 and onset stop laryngeal contrasts

Task 1 tests the effect of F0 on perception of onset stop voicing. Based on previous work (e.g., Haggard et al., 1970), it is expected that listeners will be more likely to identify an onset stop as voiceless and aspirated when the following F0 is higher.

Table 7 presents the summary of a mixed effects logistic regression model for responses of the voiced option (e.g., *best* rather than *pest*) for the data subsetted to only include participants with at least 85% accuracy identifying unambiguous consonants. The fixed effects were Pitch (High, Low), Huggins result (Fail, Pass), Dichotic Loudness result (Fail, Pass), the interaction between Pitch and Huggins, and the interaction between Pitch and Loudness. There was a random intercept for participant and for word pair. Random slopes were not included because they were highly correlated with the corresponding intercepts and prevented the model from converging.

	β	**SE**	**z value**	***p* value**
(Intercept)	–0.508	0.513	–0.989	0.322
Pitch Low	0.593	0.194	3.06	0.00225
Huggins Pass	–0.343	0.356	–0.964	0.335
Loudness Pass	–0.0877	0.353	–0.249	0.804
Pitch Low * Huggins Pass	0.525	0.326	1.61	0.107
Pitch Low * Loudness Pass	–0.471	0.323	–1.46	0.145

Table 7: Regression model for voiced responses, Task 1. *Reference Levels: Pitch = High, Huggins = Fail, Loudness = Fail.*

The expected effect was observed regardless of participants' performance on the headphone checks; listeners were more likely to perceive an onset stop as voiced (unaspirated) when the following vowel had a lower F0. Note that because of the existence of the interaction terms, the main effect of Pitch is providing the estimate for listeners who failed both headphone checks (even though the interactions are not significant).

There were no significant effects of either headphone check, nor significant interactions between the headphone checks and the effect of pitch.

These results suggest that headphone checks are not providing any benefit beyond the filtering that is already accomplished by excluding participants with below 85% accuracy identifying clear items.

Figure 2 shows random slopes for pitch by participant, in order to visualize potential effects of the headphone checks and evaluate whether variation across participants was decreased by the headphone checks. These slopes come from a model with random slopes by participant and no random intercept by word pair, which excluded headphone checks as factors, both because a model would not converge that included these factors along with random slopes and also so that any effects of headphone checks would be reflected in the by-participant slopes rather than being captured by fixed effects. While the effect of pitch was slightly higher for participants who passed the headphone checks, there was little difference in the degree of variation across participants. Variance tests showed no evidence for a difference either based on the dichotic loudness test ($F = 0.692$, $p = 0.181$) or the Huggins test ($F = 0.735$, $p = 0.271$).

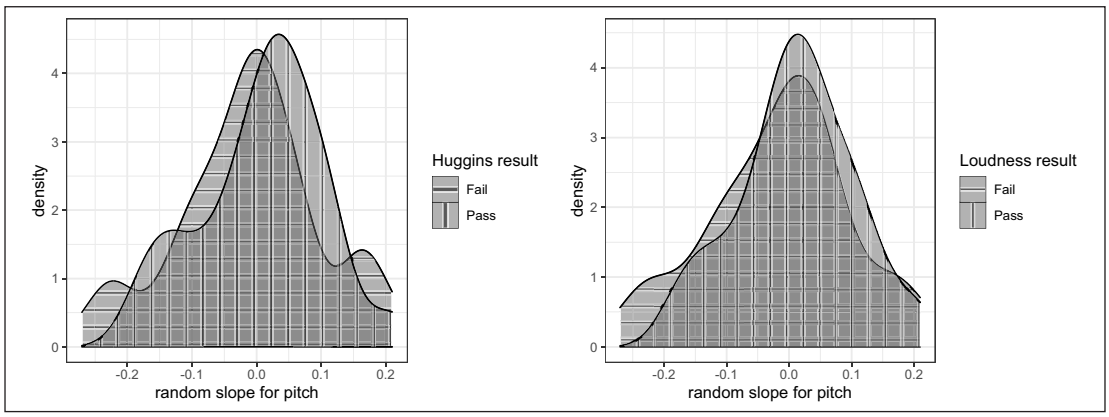

Figure 2: Random slopes for the effect of pitch on voiced responses by participant, from a model that excluded headphone checks as factors. Participants are grouped by their results on the Huggins check (left) and the dichotic loudness check (right).

Since one of the main goals in collecting usable data is to exclude participants who are not paying attention to the task, we might also examine the effect of participants' approval rate on Prolific. Additionally, if phones are predictive of low attention or poor audio, instructing listeners that the task should only be done on a computer or tablet should increase the size of the effect of pitch.

Table 8 presents the summary of a mixed effects logistic regression model for responses of the voiced option (e.g., *best* rather than *pest*) for the full data, with no subsetting based on accuracy for clear items. Using data already subsetted based on accuracy might obscure the

effects of these Prolific settings in excluding participants who are inattentive or unable to hear the stimuli. This model is run separately from the preceding model both in order to include the full data and also because there is substantial collinearity between the headphone checks and device restriction instructions as predictors, so a model including both does not converge. The fixed effects were Pitch (High, Low), Device Restrictions (None, NoPhone), Approval Rate (90, 95), the interaction between Pitch and Device Restrictions, and the interaction between Pitch and Approval Rate. There was a random intercept for participant and for word pair.

There were no significant effects of either participants' approval rate on Prolific or the instructions about which devices should be used for the task.

	β	**SE**	**z value**	***p* value**
(Intercept)	–0.286	0.511	–0.561	0.575
Pitch Low	0.583	0.253	2.3	0.0213
DeviceRestrictions None	–0.319	0.328	–0.972	0.331
ApprovalRate 90	–0.211	0.331	–0.64	0.522
Pitch Low * DeviceRestrictions None	–0.0989	0.309	–0.32	0.749
Pitch Low * ApprovalRate 90	–0.052	0.311	–0.167	0.867

Table 8: Regression model for voiced responses, Task 1. *Reference Levels: Pitch = High, DeviceRestrictions = NoPhone, ApprovalRate = 95.*

3.4. Task 2: Vowel height category boundaries

Task 2 tests the effect of exposure to shifted F1 on subsequent vowel category boundaries. Based on previous work (e.g., Ladefoged & Broadbent, 1957), listeners should be more likely to identify ambiguous items on /ɛ-ɪ/ continua as being /ɛ/ after exposure to mid vowels with a lowered F1 than after exposure to mid vowels with a raised F1. The F1 of the test item should also predict responses; listeners should be more likely to perceive a vowel as /ɛ/ if it has a higher F1.

Table 9 presents the summary of a mixed effects logistic regression model for responses of the lower vowel (e.g., *pet* rather than *pit*) for the data subsetted to only include participants with at least 85% accuracy identifying unambiguous consonants. The fixed effects were F1 step, Exposure Condition (Raised F1, Lowered F1), Huggins result (Fail, Pass), Dichotic Loudness result (Fail, Pass), the interaction between F1 step and Huggins, the interaction between Exposure and Huggins, the interaction between F1 step and Loudness, and the interaction between Exposure and Loudness. There was a random intercept for participant and for word pair. Random slopes were not included because they were highly correlated with the corresponding intercepts and prevented the model from converging.

	β	SE	z value	*p* value
(Intercept)	–5.06	0.416	–12.2	< 0.0001
Exposure RaisedF1	–1.68	0.306	–5.5	< 0.0001
F1step	1.79	0.104	17.3	< 0.0001
Huggins Pass	–1.09	0.599	–1.82	0.0683
Loudness Pass	0.217	0.576	0.377	0.706
Exposure RaisedF1 * Huggins Pass	0.0804	0.518	0.155	0.877
F1step * Huggins Pass	0.265	0.176	1.51	0.132
Exposure RaisedF1 * LoudnessPass	0.0748	0.512	0.146	0.884
F1step * LoudnessPass	–0.0807	0.17	–0.475	0.635

Table 9: Regression model for responses of the lower vowel, Task 2. *Reference Levels: Exposure Condition = Lowered F1, Huggins = Fail, Loudness = Fail.*

The expected effects were observed regardless of participants' performance on the headphone checks. A higher F1 made listeners more likely to identify a vowel as being mid rather than high. Prior exposure to mid vowels with a raised F1 (versus lowered F1) resulted in fewer identifications of the testing stimuli as having mid vowels: The category boundary had a higher F1 in the raised F1 exposure condition than in the lowered F1 exposure condition. As before, note that the main effects are for participants who failed both headphone checks, given the presence of the interaction terms.

There was a marginal main effect of the Huggins check; listeners who passed it were more likely to identify vowels as being higher (i.e. /ɪ/ rather than /ɛ/). This might suggest that the Huggins check results in a population with a different category boundary. The demographic patterns described above show regional differences based on the headphone check results; people who pass the Huggins check are more likely to be from the west or northeast, and less likely to be from the southeast or midwest.

There were no significant interactions between either headphone check and either of the experimental variables.

These results suggest that headphone checks are not providing any benefit beyond the filtering that is already accomplished by excluding participants with below 85% accuracy identifying clear items.

Figures 3 and **4** show random slopes by participant for the exposure condition and F1 step, respectively. These figures visualize the potential effects of headphone checks. The slopes come from a model with random slopes by participant and no random intercept by word pair, which excluded headphone checks as factors. Both factors exhibited little difference in the degree of variation across participants based on headphone check results. Variance tests confirmed the visual patterns; there was no evidence for a difference in exposure condition variance based on dichotic

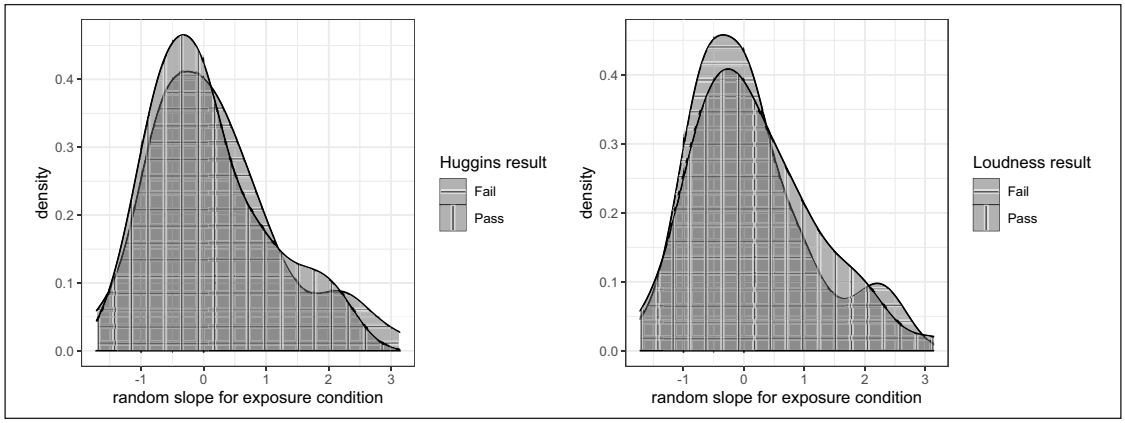

Figure 3: Random slopes for the effect of exposure condition on vowel identifications by participant, from a model that excluded headphone checks as factors. Participants are grouped by their results on the Huggins check (left) and the dichotic loudness check (right).

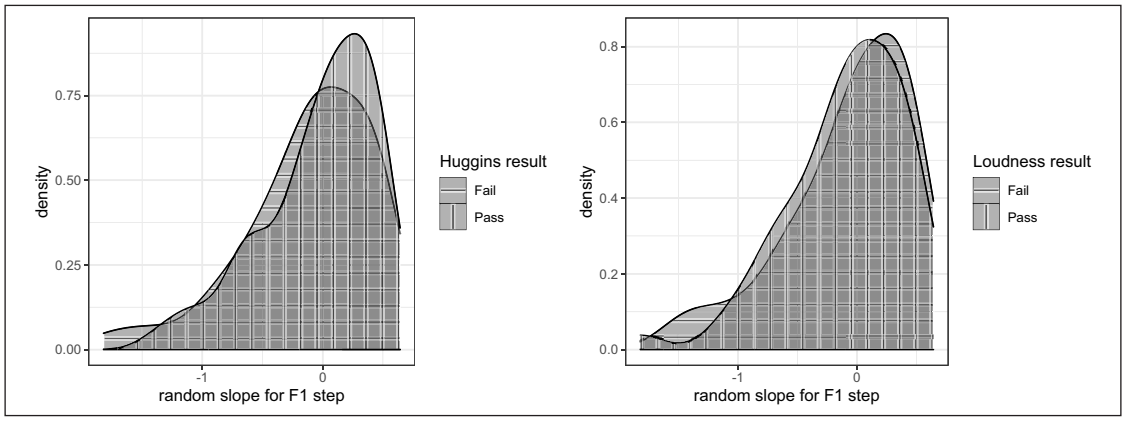

Figure 4: Random slopes for the effect of F1 step on vowel identifications by participant, from a model that excluded headphone checks as factors. Participants are grouped by their results on the Huggins check (left) and the dichotic loudness check (right).

loudness ($F = 1.05$, $p = 0.851$) or Huggins ($F = 0.775$, $p = 0.363$), nor a difference in F1 step variance based on dichotic loudness ($F = 0.89$, $p = 0.671$) or Huggins ($F = 0.759$, $p = 0.325$).

The next set of factors to be examined are participants' approval rate on Prolific and instructions about which devices the task could be done on. **Table 10** presents the summary of a mixed effects logistic regression model for responses of the lower vowel (e.g., *pet* rather than *pit*) for the full data. The fixed effects were F1 step, Exposure (Raised F1, Lowered F1), Device Restrictions (None, NoPhone), Approval Rate (90, 95), the interaction between F1 step and Device Restrictions, the interaction between Exposure and Device Restrictions, the interaction between F1 step and Approval Rate, and the interaction between Exposure and Approval Rate. There was a random intercept for participant and for word pair.

	β	SE	z value	*p* value
(Intercept)	–5.84	0.56	–10.4	< 0.0001
Exposure RaisedF1	–2.04	0.414	–4.92	< 0.0001
F1step	1.94	0.156	12.5	< 0.0001
DeviceRestrictions None	1.67	0.599	2.78	0.00534
ApprovalRate 90	–0.791	0.53	–1.49	0.136
Exposure RaisedF1 * DeviceRestrictions None	0.665	0.496	1.34	0.18
F1step * DeviceRestrictions None	–0.479	0.174	–2.75	0.00587
Exposure RaisedF1 * ApprovalRate 90	0.13	0.48	0.271	0.786
F1step * ApprovalRate 90	0.309	0.152	2.03	0.0422

Table 10: Regression model for responses of the lower vowel, Task 2. *Reference Levels: Exposure Condition = Lowered F1, DeviceRestrictions = NoPhone, ApprovalRate = 95.*

There were more responses of /ɛ/ among people with no instructions about device restrictions. This could potentially be an effect of the frequency response produced by phones. Measurements of formants are biased towards harmonics (Chen, Whalen, & Shadle, 2019). Patterns of intensity might similarly affect perception; for example, Lotto, Holt, and Kluender (1997) find that increasing the intensity of lower frequencies increases identifications of vowels as tense. Alternatively, this could be an effect of the dialects of the speakers who were using phones. Only 8% of northeastern participants and 15% of southwestern participants reported using phones, while 29% of midwestern participants, 24% of southeastern participants, and 25% of western participants reported using phones.

There was a smaller effect of F1 step among participants with no device restriction instructions: F1 has less of an impact on what vowel they identify the stimulus as having. This effect is consistent with participants on their phones being less able to hear the stimuli or paying less attention.

However, there was also a larger effect of F1 step among participants recruited with a lower restriction on approval rate. The cause for this small but significant effect is not clear. It might be due to the incomplete balancing of conditions, as only people in the condition with no device restrictions had the 90% approval rate threshold. However, it is worth considering that the approval rate threshold might capture something about the way that participants approach tasks that is not simply low attention; they might be less attentive to the characteristics that are being used for rejections, but might be more attentive to some aspects of the task.

3.5. Task 3: Spectral tilt and perceived vowel duration

Task 3 tests the effect of spectral tilt on perceived vowel duration. Based on previous work (Sanker, 2020), listeners should be more likely to identify vowels as long when they have a lower spectral tilt. Actual duration should also be a predictor of perceived duration, and listeners should compensate for the voicing of the coda, more often identifying vowels as long when they are presented with a voiceless coda.

Table 11 presents the summary of a mixed effects logistic regression model for 'long' responses for the data subsetted to only include participants with at least 85% accuracy identifying unambiguous consonants. The fixed effect were Duration Step, Original Coda (p, b), Spliced Coda (p, b), Spectral Tilt (high, low), Huggins result (Fail, Pass), Dichotic Loudness result (Fail, Pass), the interaction between Duration Step and Huggins, the interaction between Original Coda and Huggins, the interaction between Spliced Coda and Huggins, the interaction between Spectral Tilt and Huggins, the interaction between Duration Step and Loudness, the interaction between Original Coda and Loudness, the interaction between Spliced Coda and Loudness, and the interaction between Spectral Tilt and Loudness. There was a random intercept for participant, and a random slope for Duration Step, Original Coda, Spliced Coda, and Spectral Tilt by participant. Note that unlike the previous two studies, including these random slopes was possible, because the correlation between the intercept and the random slopes was low.

	β	**SE**	**z value**	***p* value**
(Intercept)	–2.8	0.325	–8.63	<0.0001
DurationStep	0.894	0.0942	9.49	<0.0001
OrigCoda p	–0.448	0.138	–3.25	0.00114
SplicedCoda p	0.603	0.164	3.68	0.000236
Tilt Low	0.154	0.246	0.627	0.531
Huggins Pass	–1.02	0.542	–1.89	0.0591
Loudness Pass	0.793	0.535	1.48	0.138
DurationStep * Huggins Pass	0.222	0.159	1.39	0.164
OrigCoda p * Huggins Pass	0.233	0.233	1.0	0.317
SplicedCoda p * Huggins Pass	–0.633	0.276	–2.29	0.0219
Tilt Low * Huggins Pass	1.11	0.412	2.7	0.00695
DurationStep * Loudness Pass	0.01	0.157	0.064	0.949
OrigCoda p * Loudness Pass	–0.0123	0.23	–0.054	0.957
SplicedCoda p * Loudness Pass	–0.513	0.273	–1.88	0.0604
Tilt Low * Loudness Pass	–0.256	0.408	–0.627	0.531

Table 11: Regression model for 'long' responses, Task 3. *Reference Levels: OrigCoda = b, SplicedCoda = b, Tilt = High, Huggins = Fail, Loudness = Fail.*

Most of the same effects found by Sanker (2020) were observed regardless of participants' performance on the headphone checks. Listeners were more likely to identify vowels with longer duration as being long, less likely to identify vowels as long when they had originally been produced with a voiceless coda, and more likely to identify vowels as long when they were presented with a voiceless coda.

As before, note that the main effects are for participants who failed both headphone checks, given the presence of the interaction terms. Among participants who failed both headphone checks, there was not a significant effect of spectral tilt in predicting identifications of the vowel as long. It is perhaps worthwhile to note that in a model that omits headphone checks as factors, listeners were significantly more likely to identify vowels as long if they had lower spectral tilt: $\beta = 0.509$, SE $= 0.175$, z $= 2.91$, p $= 0.00364$.

There was a marginally significant main effect of the Huggins check: Listeners who passed the Huggins check were less likely to identify vowels as being long. There was a trend of the Dichotic Loudness check in the opposite direction, increasing the odds that a vowel would be identified as long. This might be an issue resulting from the moderate correlation of Huggins and Dichotic Loudness. If these results are indeed capturing something real, it is unclear what is driving them.

The interaction between Spliced Coda and Huggins is significant; the effect of the spliced coda was eliminated among participants who passed the Huggins check. This might indicate that the Huggins check has a relationship with how participants are approaching the task; if listeners categorize vowel duration before hearing the coda, they will not exhibit an effect of coda characteristics. The median time spent on the task was almost 20% more for participants who failed the Huggins check than for participants who passed.

The interaction between Spectral Tilt and Huggins is significant; listeners who passed the Huggins check were more likely to identify vowels with lower spectral tilt as being longer (the effect predicted by previous work).

The interaction between Spliced Coda and Dichotic Loudness is marginally significant; the effect of the spliced coda was eliminated among participants who passed the dichotic loudness check. Similar to the interaction between Spliced Coda and Huggins, this likely reflects time spent on the task: Participants who failed the dichotic loudness check spent 9% longer on the task than participants who passed.

Figures 5, 6, 7, and **8** show random slopes by participant for duration step, original coda, spliced coda, and spectral tilt, respectively. These figures visualize the potential effects of headphone checks. These slopes come from a model that excluded headphone checks as factors. While the center of the distribution differs based on the headphone checks for the effects of spliced coda and spectral tilt (consistent with the model results), all of the factors exhibited relatively little difference

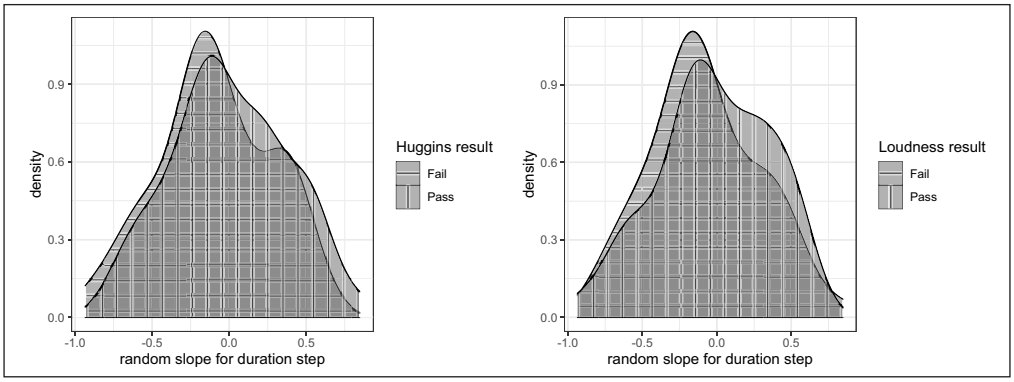

Figure 5: Random slopes for the effect of duration step on duration categorizations by participant, from a model that excluded headphone checks as factors. Participants are grouped by their results on the Huggins check (left) and the dichotic loudness check (right).

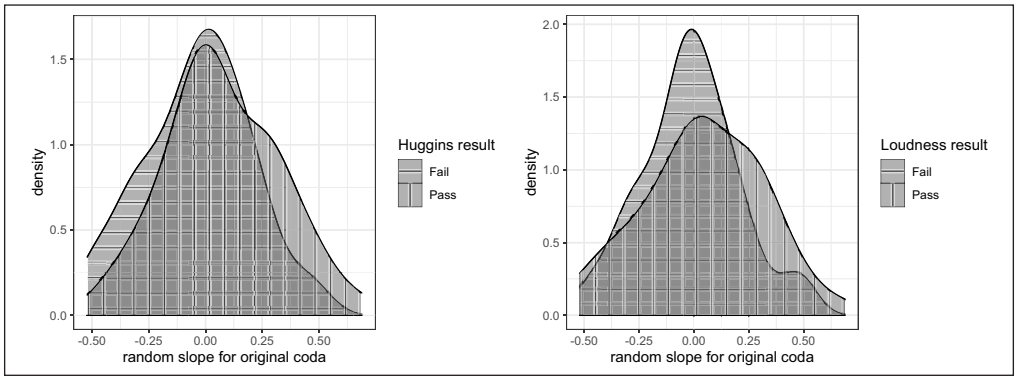

Figure 6: Random slopes for the effect of original coda voicing on duration categorizations by participant, from a model that excluded headphone checks as factors. Participants are grouped by their results on the Huggins check (left) and the dichotic loudness check (right).

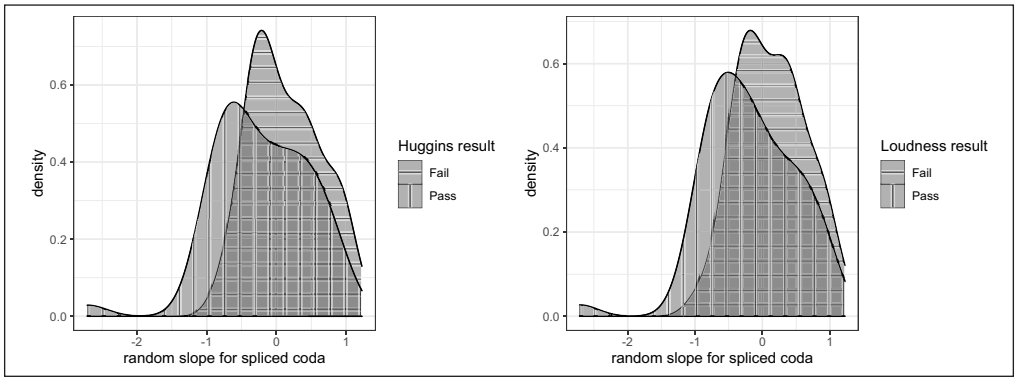

Figure 7: Random slopes for the effect of spliced coda voicing on duration categorizations by participant, from a model that excluded headphone checks as factors. Participants are grouped by their results on the Huggins check (left) and the dichotic loudness check (right).

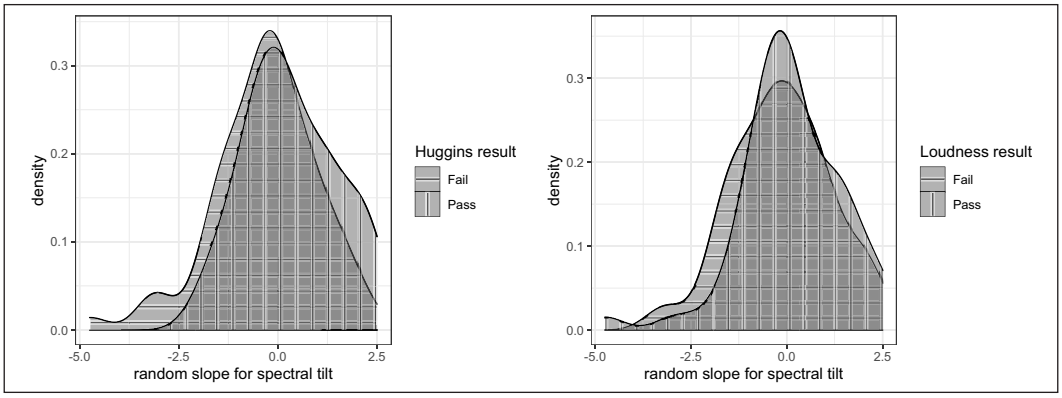

Figure 8: Random slopes for the effect of spectral tilt on duration categorizations by participant, from a model that excluded headphone checks as factors. Participants are grouped by their results on the Huggins check (left) and the dichotic loudness check (right).

in the degree of variation across participants based on headphone check results. Variance tests confirmed the visual patterns; there was no evidence for a difference in duration step variance based on dichotic loudness ($F = 1.04$, $p = 0.89$) or Huggins ($F = 1.05$, $p = 0.86$), nor a difference in original coda variance based on dichotic loudness ($F = 1.53$, $p = 0.118$) or Huggins ($F = 1.12$, $p = 0.675$), nor an effect of spectral tilt variance based on dichotic loudness ($F = 1.09$, $p = 0.739$) or Huggins ($F = 0.779$, $p = 0.375$). However, there was evidence for a small difference in spliced coda variance based on dichotic loudness ($F = 1.82$, $p = 0.0286$) and Huggins ($F = 1.86$, $p = 0.0218$); in both cases, participants who passed the headphone check are found to have more variance.

The next set of factors to be examined are participants' approval rate on Prolific and the given instructions about which devices the task could be done on. **Table 12** presents the summary of a mixed effects logistic regression model for 'long' responses for the full data. The fixed effects were Duration Step, Original Coda (p, b), Spliced Coda (p, b), Spectral Tilt (high, low), Device Restrictions (None, NoPhone), Approval Rate (90, 95), the interaction between Duration Step and Approval Rate, the interaction between Original Coda and Approval Rate, the interaction between Spliced Coda and Approval Rate, the interaction between Spectral Tilt and Approval Rate, the interaction between Duration Step and Device Restrictions, the interaction between Original Coda and Device Restrictions, the interaction between Spliced Coda and Device Restrictions, and the interaction between Spectral Tilt and Device Restrictions. There was a random intercept for participant, and a random slope for Duration Step, Original Coda, Spliced Coda, and Spectral Tilt by participant.

The interaction between Duration Step and Approval Rate was marginally significant. Participants recruited with the lower threshold for approval rate had a slightly larger effect of duration step: Actual duration was a larger predictor of whether the vowel was identified as being long.

	β	SE	z value	*p* value
(Intercept)	–3.2	0.455	–7.04	< 0.0001
DurationStep	0.957	0.132	7.25	< 0.0001
OrigCoda p	–0.29	0.183	–1.59	0.113
SplicedCoda p	–0.467	0.227	–2.05	0.04
Tilt Low	0.924	0.328	2.82	0.00482
DeviceRestrictions None	0.863	0.543	1.59	0.112
ApprovalRate 90	–0.753	0.529	–1.42	0.155
DurationStep * DeviceRestrictions None	–0.141	0.158	–0.893	0.372
OrigCoda p * DeviceRestrictions None	–0.114	0.219	–0.522	0.602
SplicedCoda p * DeviceRestrictions None	0.643	0.274	2.35	0.0188
Tilt Low * DeviceRestrictions None	–0.545	0.397	–1.37	0.17
DurationStep * ApprovalRate 90	0.304	0.155	1.96	0.0504
OrigCoda p * ApprovalRate 90	0.103	0.209	0.492	0.623
SplicedCoda p * ApprovalRate 90	0.198	0.265	0.748	0.455
Tilt Low * ApprovalRate 90	–0.091	0.389	–0.234	0.815

Table 12: Regression model for 'long' responses, Task 3. *Reference Levels: OrigCoda = b, SplicedCoda = b, Tilt = High, DeviceRestrictions = NoPhone, ApprovalRate = 95.*

The main effect of Spliced Coda (i.e., for participants in the NoPhone condition with the 95% approval rate threshold) is actually the opposite of what was predicted and what is observed in the previous model; participants were less likely to identify vowels as long when they were presented with a voiceless coda, though the effect is relatively small. This might suggest a different task strategy, in which listeners categorize vowel duration in a way that aligns with what is expected based on the coda, rather than compensating for what is expected based on the coda. The two effects of spliced coda voicing might reflect different stages of processing; listeners in the NoPhone condition completed the task more quickly than listeners with no device restrictions.

The interaction between Spliced Coda and Device Restrictions was significant. The effect of the spliced coda is eliminated when participants were given no instructions about what devices could be used.

4. Discussion

The results help demonstrate how headphone checks are functioning, what types of studies they are likely to affect, and why those effects arise. While most phonological and phonetic patterns do not exhibit clear effects of headphone checks, there are some significant effects. In

addition to effects of headphone checks that can be explained by acoustic characteristics of the audio produced by headphones, there are some effects that seem to be due to patterns in the demographic characteristics of which participants pass or fail headphone checks. Both types of effects have implications for how experimental design might impact the interpretability and generalizability of results, based on what acoustic characteristics, task strategies, or demographic patterns the examined phenomena are sensitive to.

4.1. Interpretation of results

The results of the perception tasks in this study show limited benefits of headphone checks. For a range of factors, there was no evidence that listeners who passed the headphone checks (Huggins, Milne et al., 2021 and dichotic loudness, Woods et al., 2017) produced different results than listeners who failed these checks. These tests included F0 as a cue to onset voicing, F1 as a cue to vowel height, effect of exposure to manipulated F1 on subsequent category boundaries, and actual vowel duration as a predictor of perceived vowel duration. Headphone checks do have some significant or marginal main effects and interactions with phonological factors; however, it is likely that the effects arise in several different ways.

One factor which was substantially impacted by headphone checks was spectral tilt as a predictor of perceived vowel duration, in Task 3. Participants who passed the Huggins check had a larger effect of spectral tilt on perceived vowel duration, replicating the previously observed effect; participants who failed the Huggins check did not exhibit a significant effect. This expected effect depends on the relationship between frequency and perceived loudness and between loudness and perceived duration; perceived intensity increases with frequency, so a sound with lower spectral tilt is likely to be perceived as louder and subsequently longer (Sanker, 2020). Intensity at different frequencies can vary substantially across devices, which can obscure effects that depend on spectral tilt. The reason why the Huggins check results in a stronger effect of spectral tilt on perceived vowel duration may be that the check is specifically selecting for high-quality headphones, which are better at producing low frequencies and thus will capture the differences in spectral tilt, whereas devices which attenuate low frequencies reduce the difference between the spectral tilt manipulations. The Huggins check is likely to similarly provide a benefit for capturing other effects that also depend on intensity relative to frequency.

Not all effects of headphone checks seems to reflect an improvement in capturing expected effects. The effect of spliced coda voicing on perceived vowel duration in Task 3 was only significant among participants who failed the headphone checks, and was eliminated among participants who passed one or both checks. The by-participant variation was also larger among participants who passed the headphone checks. This might be related to how participants approached the tasks; participants who passed headphone checks completed the study more quickly than participants who failed them. Given the faster responses, it is possible that the effect

of coda voicing was absent among these participants because they often responded before hearing the codas. One of the reasons why participants who pass headphone checks might complete the tasks more quickly is that they have more experience in experimental tasks. Participants who have more experience with perception experiments might also be more likely to recognize headphone checks, which could make them less likely to misinterpret the instructions and more likely to know what listening setup will allow them to pass the check. Listeners who have done the headphone check tasks before may also have higher accuracy in them due to practice; Akeroyd, Moore, and Moore (2001) find a significant improvement in identification of melodies produced by Huggins dichotic pitch stimuli from the first block of items to the second block. Another possible reason for the speed difference is that the headphone checks reduce the average age of participants, and younger participants complete the task more quickly ($r(109) = 0.27$, $p = 0.00415$), but the difference in task duration based on headphone check results is larger than what would be predicted just based on the relationship between age and task duration.

Headphone checks also produce differences in participants' regional origins, which can impact the results in ways not related to audio clarity. There was a marginal main effect of Huggins on the category boundary between /ɪ/ and /ɛ/ in Task 2; listeners who passed the Huggins check were more likely to identify ambiguous vowels as being /ɪ/. Participants from the midwest and southeast are more likely to fail the Huggins check, while participants from the northeast and west have better odds of passing the Huggins check. Several known shifts might contribute to the Huggins check producing a different vowel category boundary: /ɪ/ is lowered in the Northern California vowel shift (Eckert, 2008) and raised in the Southern vowel shift (Labov, Ash, & Boberg, 2006); a lower prototypical /ɪ/ would result in more stimuli in the mid to high range being identified as /ɪ/, while a higher prototypical /ɪ/ would result in fewer stimuli in this range being identified as /ɪ/.

The effect of headphone checks varies depending on what is being examined; for example, perception of vowel quality based on F1 versus perception of vowel duration based on spectral tilt or coda environment. For many perceptual effects examined here, headphone checks have no apparent impact. For other patterns, headphone checks might impact results either because headphone use is directly important for capturing the pattern or because headphone checks are selecting for a population that produces different results because of their dialects or approaches to the task. The impact of headphone checks seems to depend on the nature of the particular factor being examined, rather than the robustness of the effect. None of the effects examined here are at ceiling, so there would be room for all of them to be influenced by headphone checks, but most are not. Small effects will be more susceptible to being obscured by variability, but it is not clear that headphone checks are reducing variability across listeners in relevant ways; none of the factors examined showed evidence for participants who passed the headphone checks having less variability than participants who failed them.

4.2. Implications

One benefit of the rather limited effects of headphone checks is that results across studies are likely to be comparable even though some published studies have used headphone checks and others have not. A few studies compare how their results would turn out depending on whether or not they exclude participants who failed the headphone check, and report no difference (Ringer et al., 2022; Shen & Wu, 2022). However, headphone checks can influence the results for some phonological and phonetic effects, for several potential reasons.

What participants are passing headphone checks? One part of what headphone checks are doing is selecting for headphones, as has been demonstrated previously (Milne et al., 2021; Woods et al., 2017). The data presented here shows that there are also several demographic differences between participants who pass and fail headphone checks, including age, gender, education, and geographic region. There is also evidence that the headphone checks select for participants who approach the tasks differently, in particular resulting in faster task completion. While these relationships raise some concerns, knowing more about the factors that predict variation in online studies allows us to better control for them.

Participants who pass headphone checks are more likely to be from the northeast or the west, while participants who fail the headphone checks are more likely to be from the southeast or the midwest. This can produce differences in category boundaries or even what contrasts exist. Online participants come from many different locations; this variability may be desirable in capturing broad patterns among speakers of different dialects of a language, though in other cases it may be desirable to have a more uniform sample or to recruit participants from particular dialect regions. A sample biased towards a particular location might have different results than a more regionally balanced sample, so having regional information can be important in interpreting results and evaluating whether the observed patterns are likely to be generalizable. Regional information can be collected in post-task demographic surveys, as was done in this study. It is also possible to recruit participants using eligibility filters; Prolific includes the option to recruit based on state-level location and state where the participant was born.

Participants who pass the headphone checks are more likely to be men than women. This result may be related to device usage, as women were more likely than men to complete the task using a phone. The small number of participants who were nonbinary or did not identify their gender were also more likely to fail the headphone checks than men were. Differences in participant gender might impact the results for some studies, given gender effects in some linguistic tasks (e.g., Namy, Nygaard, & Sauerteig, 2002).

Participants who pass headphone checks are likely to be younger than participants who fail them. One major source of this effect is likely to be hearing loss among older participants, which makes it more difficult to hear Huggins pitch stimuli (Santurette & Dau, 2007). Older and

younger individuals can also differ in phonological patterns based on sound changes in progress (e.g., Harrington, Kleber, & Reubold, 2008), and exhibit differences in linguistic tasks for a range of other reasons (e.g., Scharenborg & Janse, 2013; Shen & Wu, 2022). Given that many in-person studies use university undergraduates as participants (Peterson, 2001), it is possible that younger online participants will produce results more similar to previous in-person results, so it is important to consider how the gold standard for comparison should be determined.

Participants who pass headphone checks are less likely to have attended college or graduate/professional school. This might reflect differences in device choice (e.g., people who attended graduate/professional school are more likely than others to use built-in computer speakers, and high school graduates are more likely to use over-the-ear headphones while people who attended college were more likely to use earbuds/in-the-ear headphones). People who have gone further in higher education may be more familiar with experimental tasks and the motivations behind them based on coursework and exposure to research on campus, which could impact how they approach tasks. Education is also related to socio-economic status, and both can be predictors of dialect features (Baranowski, 2017; Labov, Rosenfelder, & Fruehwald, 2013).

Participants who pass the headphone checks seem to be approaching the task differently in some ways; those who passed either headphone check had faster median task completion times than participants who failed the headphone checks. Speed can impact results, based on how much of the stimulus a participant hears before making a decision and also based on the stage of processing that an effect arises in (e.g., Burton, Baum, & Blumstein, 1989). This effect might be related to familiarity with online experiments; however, the post-task questionnaire did not ask about about participation in previous tasks, so this interpretation remains speculative.

While one benefit of online studies is that we might see more individual variation due to getting a more diverse population, variation in the listening setup makes it difficult to distinguish between effects of the individual and effects of listening device, distractions, background noise, or other external factors. Not only can variation across listeners obscure potential demographic predictors, the correlation between many personal characteristics and the devices that listeners use, as demonstrated in this study and elsewhere (Haan et al., 2019; Lambert & Miller, 2015; Passell et al., 2021) could produce the appearance of individual differences or demographic predictors that are actually about devices. For example, Kopiez, Wolf, Platz, and Mons (2016) found that professional musicians and audio engineers were more accurate than other people in identifying whether a stimulus was a recording of real orchestra or created from pre-recorded orchestra sample libraries. However, their data was collected online, so it is unclear whether the result indicates that people who work with music professionally are more sensitive to subtle acoustic differences or that these individuals have higher quality audio equipment. If listeners' ability to hear the stimuli is substantially impacted by the listening setup, the variation in listening setup across individuals could produce the appearance of individual differences in perception.

However, there is no clear evidence in the current study that headphone checks reduce variation across participants.

4.3. Recommendations

Headphone checks are excluding a large number of potential participants, and for many types of studies it is not clear that these exclusions are improving the data. This unnecessarily increases the time and money spent on data collection, and may be a particular problem for studies recruiting participants from a limited population. Both of the common headphone checks exclude some participants who are wearing headphones (Milne et al., 2021; Woods et al., 2017), so rejections based on headphone checks may seem confusing and unjustified to participants. At least within Prolific, instructions about what devices are suitable for a task can also help in recruiting participants with the desired listening setup; the results of this study suggest that participants do follow these instructions.

Even for tasks in which headphones are beneficial, it is important to note that attention in one part of a study does not guarantee attention in other parts. This may become a particular issue with participants who know the function of certain tasks. Sometimes users of recruitment platforms for online paid tasks will complete a large number of similar studies, and will discuss studies with online communities of other users (Aguinis et al., 2021; Chandler et al., 2014). This means that they are often aware of how different tasks are used and may know that headphone checks are often used to exclude participants from participation. They may be familiar with headphone checks and pay more attention to the portion of a task that looks like a headphone check than to other portions of the task. They might turn off background music or other distractors during the headphone check or put on headphones to pass the headphone check, but resume their preferred listening setup for the rest of the task. Headphone checks usually occur at the beginning of a task, so they are likely to receive greater attention than other parts of the task even if participants are not aware of their function. The ideal checks for attention and audio clarity will occur throughout the task. For some tasks, clear items are naturally built-in; for example, items at each end of a continuum are likely to have a clear correct answer, as used by Brown et al. (2018); Luthra et al. (2021); Mills et al. (2022); Saltzman and Myers (2021).

Rather than headphone checks for all studies, it will often be preferable to instead have a test for ability to hear a selection of clear items similar to the target stimuli. These might be a subset of the target stimuli, or might be designed specifically to serve as an audio and attention check. These checks can cover a wide range of potential factors, such as background noise, distractions, or not being a fluent speaker of the target language. Moreover, such tests can be aimed at specifically evaluating the participant's ability to do the particular task of interest. Not all tasks will be sensitive to the same types of issues; for example, bleed between channels is unlikely to have a large impact on vowel category boundaries, but would be a serious problem

for a dichotic listening task. Headphone checks are important when the task requires different audio presented to each ear, and might also be beneficial if the stimuli depend on an acoustic characteristic that is substantially affected by device in a way that also impacts headphone check results (e.g., spectral tilt).

Data can differ based on the recruitment platform. Previous work has demonstrated that participants recruited via Prolific produce better data than participants recruited via Amazon MTurk (Peer, Brandimarte, Samat, & Acquisti, 2017; Peer et al., 2021; Uittenhove, Jeanneret, & Vergauwe, nd). Platforms also differ in the demographics of the participants, such as location, income, and how much time per week they spend completing tasks through that platform (Peer et al., 2017), which may be a consideration in choosing where to recruit.

Recruitment websites provide a range of filters, including native language, location, gender, and approval rate in previous studies. Prolific has a particularly large set of filters that can be used, though they are self-reported by participants. Demographic filters can be valuable for reaching the target population and for controlling potential confounding variables. Approval rate has been demonstrated previously to be a predictor of accuracy (Peer et al., 2014). Within the current study, participants recruited with higher approval rate thresholds had higher accuracy for clear items. However, the participants recruited at the lower approval rate actually exhibited a larger effect for some of the predictors: F1 step in predicting the identifications of vowel category, and vowel duration step in predicting the categorization of a vowel as long. This may indicate that approval rating on participant recruitment websites is a somewhat more complicated metric to interpret, particularly when comparing relatively high thresholds, as was done in this study; Peer et al. (2014) compared results using 95% as the lower threshold versus the upper threshold. Participants with slightly more rejected submissions might have lower overall attention to the tasks, but might also have different approaches to the tasks, such as failing catch trials based on expecting consistency within a task.

Headphone checks will not produce uniformity across participants. The current results show rather limited effects of headphone checks, and no reduction in variation across participants based on headphone checks. There are many sources of variation across participants in any study, and additional sources of variation for online studies, which is probably why many studies comparing online and in-person versions of the same experiment often find slightly lower accuracy or smaller effect sizes in the online version (e.g., Cooke & García Lecumberri, 2021; Elliott et al., 2022; Slote & Strand, 2016; Wolters et al., 2010; A. C. L. Yu & Lee, 2014). Even if all participants are wearing headphones, there is still likely to be variation based on the particular headphones, as well as other factors like their sound card, background noise, and distractions. An online study will usually require a larger sample size than a comparable in-person study, in order to handle the additional variability across participants, and headphone checks will not decrease the number of participants necessary for reasonably powered online studies. Large sample sizes are particularly

important in online studies looking for individual differences or effects of demographic factors, which are likely to be confounded to some degree with effects of the listening setup. Post-task surveys are useful in controlling for aspects of the participants' listening setup that may be correlated with the factors of interest, such as device usage, distractions, background noise, and attention to the task. However, controlling for some factors is still unlikely to fully account for the sources of variability in online participants' setups.

5. Conclusions

The results suggest that headphone checks often provide no benefit beyond what is accomplished by a basic attention/audio check using accuracy of identifications of clear stimuli. However, some tasks are likely to benefit from headphone checks; for example, a headphone check is likely to be useful if one of the factors of interest is the relationship between intensity and frequency. Headphone checks also exclude participants disproportionately based on several demographic factors (e.g., location, age), which has the potential to impact experimental results, based on changing the dialect features of the sample population or changing how quickly most participants complete the task. To ensure audio quality, it may be more useful to do a direct test of perceptibility along the acoustic dimensions that are likely to be relevant, though for some acoustic characteristics a headphone check may indirectly achieve the same thing.

There are many sources of variability across online participants, so ensuring headphone use might not have any measurable impact on how easily effects can be found. Increasing sample size will usually be a more reliable way to increase power than trying to reduce variability; while some methods might reduce variability, headphone checks do not seem to be providing this benefit.

Appendix

Ambiguous VOT items with F0 manipulations	
pest-best	
pet-bet	
tech-deck	
tuck-duck	
tug-dug	
tusk-dusk	
Clear consonant decision items	
bud	bug
cut	cup
pub	pug
bed	beg
peck	pep
neck	net
suck	shuck
shut	hut
thug	hug
fed	said
head	shed
theft	heft

Table 13: Words for Task 1. Note that for the ambiguous VOT items, the listed words were just the response options; the stimuli came from ambiguous productions.

Table 14: Survey of recent articles with online speech perception tasks.

[1]They note that no participants were excluded based on the headphone check.

[2]They report results from the full data, and note that excluding participants who failed the headphone check did not change results.

Article	headphone check	headphone instructions	other attention/audio checks	questions about device	questions about attention, noise, etc
Bieber & Gordon-Salant 2022	dichotic loudness		no	no	no
Brown et al 2018	dichotic loudness		accuracy on clear items	no	no
Geller et al 2021	dichotic loudness		no	no	no
Giovannone & Theodore 2021	dichotic loudness		uniformity of responses	no	no
Krumbiegel, Ufer, & Blank 2022	dichotic loudness		no	no	no
Lavan et al 2019	dichotic loudness		no	no	no
Luthra et al 2021	dichotic loudness		accuracy on clear items	no	no
McPherson, Grace, & McDermott 2022	dichotic loudness		overall performance	no	no
Mepham, Bi, & Mattys 2022	dichotic loudness		no	no	no
Merritt & Bent 2022	dichotic loudness		no	no	no
Mills et al 2022	dichotic loudness		accuracy on clear items	no	no
Nayak et al 2022	dichotic loudness[1]		catch trials	no	yes
Saltzman & Myers 2021	dichotic loudness		accuracy on clear items	no	no
Seow & Hauser 2022	dichotic loudness		catch trials	no	no
Yu 2022	dichotic loudness		no	no	no

(Contd.)

Article	headphone check	headphone instructions	other attention/audio checks	questions about device	questions about attention, noise, etc
Yu, Schertz, & Johnson 2022	dichotic loudness		no	no	no
Beier & Ferreira 2022	Huggins		accuracy on comprehension questions, reaction time	no	yes
Brekelmans et al 2022	Huggins		accuracy on clear items, catch trials	no	no
Ringer, Schroeger, & Grimm 2022	Huggins[2]		no	no	no
Tamati et al 2022	Huggins		no	no	no
Wu & Holt 2022	Huggins		no	no	no
Davidson 2020	none reported	yes	no	yes	no
Denby & Goldrick 2021	none reported	no	no	no	no
D'Onofrio 2018	none reported	no	accuracy on clear items	no	no
Getz & Toscano 2021	none reported	yes	overall performance	yes	no
Kato & Baese-Berk 2022	none reported	no	no	no	no
Manker 2020	none reported	yes	no	no	no
McHaney et al 2021	none reported	yes	no	no	no
Reinisch & Bosker 2022	none reported	yes	no	yes	no
Vujović, Ramscar, & Wonnacott 2021	none reported	no	no	no	no
Williams, Panayotov, & Kempe 2021	none reported	no	no	no	no

Competing interests

The author has no competing interests to declare.

References

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837. DOI: https://doi.org/10.1177/0149206320969787

Akeroyd, M. A., Moore, B. C., & Moore, G. A. (2001). Melody recognition using three types of dichotic-pitch stimulus. *Journal of the Acoustical Society of America*, *110*(3), 1498–1504. DOI: https://doi.org/10.1121/1.1390336

Ang, L. Y. L., Koh, Y. K., & Lee, H. P. (2017). The performance of active noise-canceling headphones in different noise environments. *Applied Acoustics*, *122*, 16–22. DOI: https://doi.org/10.1016/j.apacoust.2017.02.005

Baranowski, M. (2017). Class matters: The sociolinguistics of GOOSE and GOAT in Manchester English. *Language Variation and Change*, *29*(3), 301–339. DOI: https://doi.org/10.1017/S0954394517000217

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

Beier, E. J., & Ferreira, F. (2022). Replication of Cutler, A., & Fodor, JA (1979). Semantic focus and sentence comprehension. Cognition, 7 (1), 49–59. *Journal of Memory and Language*, *126*, Article 104339. DOI: https://doi.org/10.1016/j.jml.2022.104339

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political analysis*, *20*(3), 351–368. DOI: https://doi.org/10.1093/pan/mpr057

Bieber, R. E., & Gordon-Salant, S. (2022). Semantic context and stimulus variability independently affect rapid adaptation to non-native English speech in young adults. *Journal of the Acoustical Society of America*, *151*(1), 242–255. DOI: https://doi.org/10.1121/10.0009170

Breebaart, J. (2017). No correlation between headphone frequency response and retail price. *Journal of the Acoustical Society of America*, *141*(6), EL526–EL530. DOI: https://doi.org/10.1121/1.4984044

Brekelmans, G., Lavan, N., Saito, H., Clayards, M., & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language*, *126*, Article 104352. DOI: https://doi.org/10.1016/j.jml.2022.104352

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. F. (2018). What accounts for individual differences in susceptibility to the McGurk effect? *PloS One*, *13*(11), e0207160. DOI: https://doi.org/10.1371/journal.pone.0207160

Burton, M. W., Baum, S. R., & Blumstein, S. E. (1989). Lexical effects on the phonetic categorization of speech: The role of acoustic structure. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 567–575. DOI: https://doi.org/10.1037/0096-1523.15.3.567

Casey, L. S., Chandler, J., Levine, A. S., Proctor, A., & Strolovitch, D. Z. (2017). Intertemporal differences among MTurk workers: Time-based sample variations and implications for online data collection. *Sage Open*, *7*(2), 1–15. DOI: https://doi.org/10.1177/2158244017712774

Casler, K., Bickel, L., & Hackett, E. (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. DOI: https://doi.org/10.1016/j.chb.2013.05.009

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: Consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46*(1), 112–130. DOI: https://doi.org/10.3758/s13428-013-0365-7

Chandler, J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are impostors. *Social Psychological and Personality Science*, *8*(5), 500–508. DOI: https://doi.org/10.1177/1948550617698203

Chen, W.-R., Whalen, D. H., & Shadle, C. H. (2019). F0-induced formant measurement errors result in biased variabilities. *Journal of the Acoustical Society of America*, *145*(5), EL360–EL366. DOI: https://doi.org/10.1121/1.5103195

Clifford, S., & Jerit, J. (2014). Is there a cost to convenience? An experimental comparison of data quality in laboratory and online studies. *Journal of Experimental Political Science*, *1*(2), 120–131. DOI: https://doi.org/10.1017/xps.2014.5

Cooke, M., & García Lecumberri, M. L. (2021). How reliable are online speech intelligibility studies with known listener cohorts? *Journal of the Acoustical Society of America*, *150*(2), 1390–1401. DOI: https://doi.org/10.1121/10.0005880

Corretge, R. (2020). *Praat vocal toolkit*. http://www.praatvocaltoolkit.com.

Dandurand, F., Shultz, T. R., & Onishi, K. H. (2008). Comparing online and lab methods in a problem-solving experiment. *Behavior Research Methods*, *40*(2), 428–434. DOI: https://doi.org/10.3758/BRM.40.2.428

Davidson, L. (2020). Contributions of modal and creaky voice to the perception of habitual pitch. *Language*, *96*(1), e22–e37. DOI: https://doi.org/10.1353/lan.2020.0013

Denby, T., & Goldrick, M. (2021). The voice of experience: Causal inference in phonotactic adaptation. *Laboratory Phonology*, *12*(1), Article 5. DOI: https://doi.org/10.5334/labphon.267

Denby, T., Schecter, J., Arn, S., Dimov, S., & Goldrick, M. (2018). Contextual variability and exemplar strength in phonotactic learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(2), 280. DOI: https://doi.org/10.1037/xlm0000465

D'Onofrio, A. (2018). Personae and phonetic detail in sociolinguistic signs. *Language in Society*, *47*(4), 513–539. DOI: https://doi.org/10.1017/S0047404518000581

Eckert, P. (2008). Where do ethnolects stop? *International Journal of Bilingualism*, *12*(1-2), 25–42. DOI: https://doi.org/10.1177/13670069080120010301

Elliott, E. M., Bell, R., Gorin, S., Robinson, N., & Marsh, J. E. (2022). Auditory distraction can be studied online! A direct comparison between in-person and online experimentation. *Journal of Cognitive Psychology*, *34*(3), 307–324. DOI: https://doi.org/10.1080/20445911.2021.2021924

Geller, J., Holmes, A., Schwalje, A., Berger, J. I., Gander, P. E., Choi, I., & McMurray, B. (2021). Validation of the Iowa test of consonant perception. *Journal of the Acoustical Society of America*, *150*(3), 2131–2153. DOI: https://doi.org/10.1121/10.0006246

Getz, L. M., & Toscano, J. C. (2021). Rethinking the McGurk effect as a perceptual illusion. *Attention, Perception, & Psychophysics*, *83*(6), 2583–2598. DOI: https://doi.org/10.3758/s13414-021-02265-6

Giovannone, N., & Theodore, R. M. (2021). Individual differences in lexical contributions to speech perception. *Journal of Speech, Language, and Hearing Research*, *64*(3), 707–724. DOI: https://doi.org/10.1044/2020_JSLHR-20-00283

Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of Mechanical Turk samples. *Journal of Behavioral Decision Making*, *26*(3), 213–224. DOI: https://doi.org/10.1002/bdm.1753

Haan, M., Lugtig, P., & Toepoel, V. (2019). Can we predict device use? an investigation into mobile device use in surveys. *International Journal of Social Research Methodology*, *22*(5), 517–531. DOI: https://doi.org/10.1080/13645579.2019.1593340

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America*, *47*(2B), 613–617. DOI: https://doi.org/10.1121/1.1911936

Harrington, J., Kleber, F., & Reubold, U. (2008). Compensation for coarticulation, /u/-fronting, and sound change in standard southern British: An acoustic and perceptual study. *Journal of the Acoustical Society of America*, *123*(5), 2825–2835. DOI: https://doi.org/10.1121/1.2897042

Hauser, D., Paolacci, G., & Chandler, J. (2019). Common concerns with MTurk as a participant pool: Evidence and solutions. In F. R. Kardes, P. M. Herr & N. Schwarz (Eds.), *Handbook of research methods in consumer psychology* (pp. 319–337). New York: Routledge/Taylor & Francis Group. DOI: https://doi.org/10.4324/9781351137713-17

Hay, J., Drager, K., & Warren, P. (2009). Careful who you talk to: An effect of experimenter identity on the production of the NEAR/SQUARE merger in New Zealand English. *Australian Journal of Linguistics*, *29*(2), 269–285. DOI: https://doi.org/10.1080/07268600902823128

Kato, M., & Baese-Berk, M. M. (2022). Perceptual consequences of native and non-native clear speech. *Journal of the Acoustical Society of America*, *151*(2), 1246–1258. DOI: https://doi.org/10.1121/10.0009403

Kopiez, R., Wolf, A., Platz, F., & Mons, J. (2016). Replacing the orchestra?–The discernibility of sample library and live orchestra sounds. *PloS One*, *11*(7), Article e0158324. DOI: https://doi.org/10.1371/journal.pone.0158324

Krumbiegel, J., Ufer, C., & Blank, H. (2022). Influence of voice properties on vowel perception depends on speaker context. *Journal of the Acoustical Society of America*, *152*(2), 820–834. DOI: https://doi.org/10.1121/10.0013363

Kuznetsova, A., Bruun Brockhoff, P., & Haubo Bojesen Christensen, R. (2015). lmerTest: Tests in linear mixed effects models [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=lmerTest (R package version 2.0-29). DOI: https://doi.org/10.18637/jss.v082.i13

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English: Phonetics, phonology and sound change*. Walter de Gruyter. DOI: https://doi.org/10.1515/9783110167467

Labov, W., Rosenfelder, I., & Fruehwald, J. (2013). One hundred years of sound change in philadelphia: Linear incrementation, reversal, and reanalysis. *Language*, 30–65. DOI: https://doi.org/10.1353/lan.2013.0015

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*(1), 98–104. DOI: https://doi.org/10.1121/1.1908694

Lambert, A. D., & Miller, A. L. (2015). Living with smartphones: does completion device affect survey responses? *Research in Higher Education*, *56*(2), 166–177. DOI: https://doi.org/10.1007/s11162-014-9354-7

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition*, *193*, Article 104026. DOI: https://doi.org/10.1016/j.cognition.2019.104026

Liang, M., Zhao, F., French, D., & Zheng, Y. (2012). Characteristics of noise-canceling headphones to reduce the hearing hazard for MP3 users. *Journal of the Acoustical Society of America*, *131*(6), 4526–4534. DOI: https://doi.org/10.1121/1.4707457

Lotto, A. J., Holt, L. L., & Kluender, K. R. (1997). Effect of voice quality on perceived height of English vowels. *Phonetica*, *54*(2), 76–93. DOI: https://doi.org/10.1159/000262212

Luthra, S., Peraza-Santiago, G., Beeson, K., Saltzman, D., Crinnion, A. M., & Magnuson, J. S. (2021). Robust lexically mediated compensation for coarticulation: Christmash time is here again. *Cognitive Science*, *45*(4), e12962. DOI: https://doi.org/10.1111/cogs.12962

Manker, J. (2020). The perceptual filtering of predictable coarticulation in exemplar memory. *Laboratory Phonology*, *11*(1), Article 20. DOI: https://doi.org/10.5334/labphon.240

McAllister, T., Preston, J. L., Ochs, L., & Hitchcock, E. (2022). *Child speech perception online and in-person: measuring and managing differences.* (Talk presented at the Challenges for Change satellite workshop of LabPhon 18)

McHaney, J. R., Tessmer, R., Roark, C. L., & Chandrasekaran, B. (2021). Working memory relates to individual differences in speech category learning: Insights from computational modeling and pupillometry. *Brain and Language*, *222*, Article 105010. DOI: https://doi.org/10.1016/j.bandl.2021.105010

McPherson, M. J., Grace, R. C., & McDermott, J. H. (2022). Harmonicity aids hearing in noise. *Attention, Perception, & Psychophysics*, *84*(3), 1016–1042. DOI: https://doi.org/10.3758/s13414-021-02376-0

Mepham, A., Bi, Y., & Mattys, S. L. (2022). The time-course of linguistic interference during native and non-native speech-in-speech listening. *Journal of the Acoustical Society of America*, *152*(2), 954–969. DOI: https://doi.org/10.1121/10.0013417

Merritt, B., & Bent, T. (2022). Revisiting the acoustics of speaker gender perception: A gender expansive perspective. *Journal of the Acoustical Society of America*, *151*(1), 484–499. DOI: https://doi.org/10.1121/10.0009282

Mills, H. E., Shorey, A. E., Theodore, R. M., & Stilp, C. E. (2022). Context effects in perception of vowels differentiated by F1 are not influenced by variability in talkers'mean F1 or F3. *Journal of the Acoustical Society of America*, *152*(1), 55–66. DOI: https://doi.org/10.1121/10.0011920

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. DOI: https://doi.org/10.3758/s13428-020-01514-0

Molesworth, B. R., & Burgess, M. (2013). Improving intelligibility at a safety critical point: In flight cabin safety. *Safety Science*, *51*(1), 11–16. DOI: https://doi.org/10.1016/j.ssci.2012.06.006

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender differences in vocal accommodation: The role of perception. *Journal of Language and Social Psychology*, *21*(4), 422–432. DOI: https://doi.org/10.1177/026192702237958

Nayak, S., Gustavson, D. E., Wang, Y., Below, J. E., Gordon, R. L., & Magne, C. L. (2022). Test of prosody via syllable emphasis ("TOPsy"): Psychometric validation of a brief scalable test of lexical stress perception. *Frontiers in Neuroscience*, *16*, Article 765945. DOI: https://doi.org/10.3389/fnins.2022.765945

Olive, S., Khonsaripour, O., & Welti, T. (2018). A survey and analysis of consumer and professional headphones based on their objective and subjective performances. In *Proceedings of the 145th Audio Engineering Society Convention.*

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*(5), 411–419. DOI: https://doi.org/10.1017/S1930297500002205

Passell, E., Strong, R. W., Rutter, L. A., Kim, H., Scheuer, L., Martini, P., Grinspoon, L., & Germine, L. (2021). Cognitive test scores vary with choice of personal digital device. *Behavior Research Methods*, *53*(6), 2544–2557. DOI: https://doi.org/10.3758/s13428-021-01597-3

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153–163. DOI: https://doi.org/10.1016/j.jesp.2017.01.006

Peer, E., David, R., Andrew, G., Zak, E., & Ekaterina, D. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*. DOI: https://doi.org/10.3758/s13428-021-01694-3

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods*, *46*(4), 1023–1031. DOI: https://doi.org/10.3758/s13428-013-0434-y

Peirce, J. W. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8–13. DOI: https://doi.org/10.1016/j.jneumeth.2006.11.017

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*(3), 450–461. DOI: https://doi.org/10.1086/323732

Reinisch, E., & Bosker, H. R. (2022). Encoding speech rate in challenging listening conditions: White noise and reverberation. *Attention, Perception, & Psychophysics*, Online-First. DOI: https://doi.org/10.3758/s13414-022-02554-8

Ringer, H., Schröger, E., & Grimm, S. (2022). Perceptual learning and recognition of random acoustic patterns. *Auditory Perception & Cognition*, *5*(3–4), 259–281. DOI: https://doi.org/10.1080/25742442.2022.2082827

Saltzman, D., & Myers, E. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*, *28*(4), 1354–1364. DOI: https://doi.org/10.3758/s13423-021-01885-1

Sanker, C. (2020). A perceptual pathway for voicing-conditioned vowel duration. *Laboratory Phonology*, *11*(1), Article 18. DOI: https://doi.org/10.5334/labphon.268

Santurette, S., & Dau, T. (2007). Binaural pitch perception in normal-hearing and hearing-impaired listeners. *Hearing research*, *223*(1-2), 29–47. DOI: https://doi.org/10.1016/j.heares.2006.09.013

Sauter, M., Draschkow, D., & Mack, W. (2020). Building, hosting and recruiting: A brief introduction to running behavioral experiments online. *Brain Sciences*, *10*(4), 251. DOI: https://doi.org/10.3390/brainsci10040251

Scharenborg, O., & Janse, E. (2013). Comparing lexically guided perceptual learning in younger and older listeners. *Attention, Perception, & Psychophysics*, *75*(3), 525–536. DOI: https://doi.org/10.3758/s13414-013-0422-4

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464. DOI: https://doi.org/10.2298/PSI1004441S

Seow, T. X., & Hauser, T. U. (2022). Reliability of web-based affective auditory stimulus presentation. *Behavior Research Methods*, *54*(1), 378–392. DOI: https://doi.org/10.3758/s13428-021-01643-0

Shalool, A., Zainal, N., Gan, K. B., Umat, C., & Mukari, S. Z. M.-S. (2017). Passive noise reduction improvement by modifying the standard audiology TDH-49 headphone. *Advanced Science Letters*, *23*(2), 1320–1324. DOI: https://doi.org/10.1166/asl.2017.8394

Shapiro, D. N., Chandler, J., & Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science*, *1*(2), 213–220. DOI: https://doi.org/10.1177/2167702612469015

Shen, J., & Wu, J. (2022). Speech recognition in noise performance measured remotely versus in-laboratory from older and younger listeners. *Journal of Speech, Language, and Hearing Research*, *65*(6), 2391–2397. DOI: https://doi.org/10.1044/2022_JSLHR-21-00557

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

Tamati, T. N., Sevich, V. A., Clausing, E. M., & Moberly, A. (2022). Lexical effects on the perceived clarity of noise-vocoded speech in younger and older listeners. *Frontiers in Psychology*, *13*, Article 837644. DOI: https://doi.org/10.3389/fpsyg.2022.837644

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (nd). From lab-based to web-based behavioural research: Who you test is more important than how you test.

Vujović, M., Ramscar, M., & Wonnacott, E. (2021). Language learning as uncertainty reduction: The role of prediction error in linguistic generalization and item-learning. *Journal of Memory and Language*, *119*, Article 104231. DOI: https://doi.org/10.1016/j.jml.2021.104231

Wagner, P., Trouvain, J., & Zimmerer, F. (2015). In defense of stylistic diversity in speech research. *Journal of Phonetics*, *48*, 1–12. DOI: https://doi.org/10.1016/j.wocn.2014.11.001

Williams, G. P., Panayotov, N., & Kempe, V. (2021). Exposure to dialect variation in an artificial language prior to literacy training impairs reading of words with competing variants but does not affect decoding skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, OnlineFirst. DOI: https://doi.org/10.1037/xlm0001094

Wolters, M. K., Isaac, K. B., & Renals, S. (2010). Evaluating speech synthesis intelligibility using Amazon Mechanical Turk. In *Proceedings of the 7th speech synthesis workshop*.

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

Wu, Y. C., & Holt, L. L. (2022). Phonetic category activation predicts the direction and magnitude of perceptual adaptation to accented speech. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(9), 913–925. DOI: https://doi.org/10.1037/xhp0001037

Wycisk, Y., Kopiez, R., Bergner, J., Sander, K., Preihs, S., Peissig, J., & Platz, F. (2022). The headphone and loudspeaker test–part I: Suggestions for controlling characteristics of playback devices in internet experiments. *Behavior Research Methods*, 1–14. DOI: https://doi.org/10.3758/s13428-022-01859-8

Yentes, R. D. (2015). *Attention and data quality in online surveys: The role of survey length, progress bars, and time disclosure.* Unpublished doctoral dissertation, North Carolina State University.

Yu, A. C. L. (2022). Perceptual cue weighting is influenced by the listener's gender and subjective evaluations of the speaker: The case of English stop voicing. *Frontiers in Psychology*, *13*, Article 840291. DOI: https://doi.org/10.3389/fpsyg.2022.840291

Yu, A. C. L., & Lee, H. (2014). The stability of perceptual compensation for coarticulation within and across individuals: A cross-validation study. *Journal of the Acoustical Society of America*, *136*(1), 382–388. DOI: https://doi.org/10.1121/1.4883380

Yu, M., Schertz, J., & Johnson, E. (2022). Do I need to repeat myself? Getting to the root of the Other Accent Effect. In J. Culburtson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th annual meeting of the cognitive science society* (pp. 1546–1552).