Laboratory Phonology
Journal of the Association for Laboratory Phonology

Baese-Berk, M. M., Staggs, C., & Jaramillo, S. (2025). Learning new speech sounds in remote and in-person protocols: Benefits, drawbacks, and considerations for future research. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 16(1), pp. 1–25. DOI: https://doi.org/10.16995/labphon.16542

Open Library of Humanities

Learning new speech sounds in remote and in-person protocols: Benefits, drawbacks, and considerations for future research

Melissa M. Baese-Berk*, Department of Linguistics, University of Chicago, Chicago, IL, USA, mmbb@uchicago.edu

Cecelia Staggs, Department of Linguistics, University of Oregon, Eugene, OR, USA, cecelias@uoregon.edu

Santiago Jaramillo, Institute of Neuroscience, University of Oregon, Eugene, OR, USA, sjara@uoregon.edu

*Corresponding author.

In-laboratory training of novel speech sounds has provided significant insight into how adult language learners learn new sounds. However, this training is often costly in terms of time in lab for participants and for experimenters. Therefore, understanding whether such paradigms can be conducted successfully in a remote setting has been of great interest to the field. In this study, we present data from participants in both in-lab and remote protocols for a two-day language learning paradigm. We compare both the results of learning across the two paradigms and the logistical aspects of conducting this research. We demonstrate that both paradigms result in learning; however, while data collection is much faster for remote protocols, there are significant trade-offs to consider in terms of individual performance and attrition within this population, among other concerns. We discuss these concerns from a methodological standpoint and explore how comparisons of training modality can be informative not only for choosing appropriate methodology but also for developing a deeper understanding of acquisition of novel speech sounds.

1. Introduction

For decades, researchers have used in-lab training protocols to investigate how individuals learn new speech sounds in language (e.g., Lively et al., 1993; Logan et al., 1991; Strange & Dittman, 1984). This work has been extremely informative and has led to insights in methodological aspects of training that impact learning (e.g., high variability phonetic training, Bradlow et al., 1997, 1999; Logan et al., 1991), cognitive factors that impact learning (e.g., Perrachione et al., 2011), and the consequences for linguistic systems more broadly after learning (e.g., Kartushina et al., 2016). However, this training is also extremely costly. Typical training paradigms take place over multiple days, with some studies taking as long as 45 sessions in the lab. Indeed, multiple day training has been shown to be highly effective and perhaps even necessary for this type of learning to occur (e.g., Earle & Myers, 2015).

Given this extended time requirement for data collection, such training studies are often difficult to run. Bringing participants into a lab for multiple days has logistical challenges in terms of scheduling, and these training paradigms often require significant experimenter oversight, requiring a lab to be staffed with a number of researchers who can conduct this training.

Even before the COVID-19 crisis, there was significant interest in conducting psycholinguistic and laboratory phonology studies remotely for many reasons. Prime among them is ease of recruitment and ease of conducting the research (Schnoebelen & Kuperman, 2010). However, conducting data collection remotely has other benefits, including the ability to recruit a more diverse population of participants (Pavlick et al., 2014).

Given these potential benefits, it is tempting to explore options to conduct training studies fully remotely. However, it is not yet known how performance on these tasks may differ in remote and in-lab protocols. For example, because attention has been shown as a critical component of laboratory-based training protocols (e.g., Mora & Mora-Plaza, 2019), it is possible that participants completing remote protocols have too many distractions in their environment to result in the same type of learning that we typically see in the laboratory. Therefore, it is critical to compare in-laboratory training to remote data collection to ensure the two are comparable.

Below, we present a summary of previous findings from in-laboratory training for novel speech sounds, focusing especially on areas where we believe in-laboratory and remote data collection may differ from one another. We then provide a summary of remote data collection in laboratory phonology, focusing first on data collected before March of 2020 (the onset of the COVID-19 crisis in the United States) and then on data collected remotely during the pandemic phase.

1.1. In-laboratory training for novel speech sounds

In-laboratory training has been used to demonstrate acquisition of very difficult sound contrasts in a learner’s second language (L2). For example, in a series of landmark papers, Logan and colleagues examined learning of English /ɹ/ and /l/ by native Japanese speakers (Bradlow et al., 1997, 1999; Lively et al., 1993; Logan et al., 1991). In these studies, they specifically examined how learners both perceive and produce the /ɹ/ /l/ contrast after significant amounts of training. This work was particularly groundbreaking because non-native contrasts had previously been thought to be extremely challenging to acquire in adulthood, and perhaps impervious to learning, even after significant exposure. Further, there was a concern that decontextualized learning in the laboratory may be inferior to classroom training and thus may not be amenable to investigating the processes underlying learning of novel speech sounds. In this set of studies, they not only demonstrated robust learning, but also elucidated some possible individual factors that impact learning. This work was critical for demonstrating that learning of novel speech sounds is possible in the laboratory, and that the laboratory is a valid testing ground for investigating myriad factors that may impact learning. Since these groundbreaking studies, researchers have investigated various aspects of training, including the amount of variability presented to participants during training (Iverson et al., 2005; Logan et al., 1991). In general, these aspects of training are a property of the stimuli and could remain stable in remote data collection, therefore we do not consider these studies further here.

Similarly, many linguistic factors have been shown to impact learning in these paradigms. Specifically, the relationship between a first and second language's phonological system has been demonstrated to be critically important to the ease or difficulty of learning new speech sounds (Best & Tyler, 2007; Flege & Bohn, 2021). While these factors are also relatively easy to control in remote data collection, it is important to note that recruitment in remote protocols often requires participants to self-report about their language background. Many in-laboratory studies also use self-reporting for language background information, but it may be more difficult to verify this information in remote protocols that typically do not have a live experimenter observing the session. That is, it is possible a researcher in an in-person study could flag a participant as possibly not meeting the language background requirements, and this may not be possible in remote settings. Further, in remote protocols, participants have been shown to give incorrect information about themselves to access more studies (e.g., Chandler & Paolacci, 2017; Aguinis et al., 2021).

Other aspects that have been shown to influence learning of novel speech sounds include a variety of cognitive factors, which are typically described under the umbrella of "individual differences." These properties are typically thought of as being relatively static. However, it is also known that some of these properties are impacted by external factors. For example, working memory has been shown to impact learning of L2 sounds (Darcy et al., 2015). However, working memory is not a static property within an individual, as it is impacted by distraction (e.g., West, 1999). Further, working memory and stress states have a reciprocal relationship (e.g., Matthews & Campbell, 2009), suggesting that working memory may vary substantially across testing environments.

Similarly, it is relatively easy to control for issues of attention and cognitive load during in-laboratory studies. In general, the circumstances for all participants in a given in-laboratory study are similar. That is, all participants in a particular study are usually tested in the same (or very similar) environments with limited distractions and are presumed to be primarily attending to the target task. However, this is not something that can be controlled in remote environments. While one participant may complete the experiment in a quiet, private room, others may complete the experiment in a noisy (both auditorily and visually) public space where there are competing demands on their attention. Because attentional control has been demonstrated to impact L2 speech sound learning (e.g., Mora & Darcy, 2023), it is possible these differences in attention and distractions during testing could impact participants differently in ways that are difficult to assess or predict in remote testing environments. While cognitive load has not been studied in as much detail in L2 speech sound learning (cf. Baese-Berk & Samuel, 2016), it is possible that distractions in the environment could impact a learners' ability to acquire novel speech sounds in remote testing environments by increasing their cognitive load during learning.

In-laboratory training and testing is often limited in terms of the participants available to a researcher. That is, in much of the research in this area, investigations are limited to populations which are geographically close to an experimenter's laboratory. This has resulted in samples from "WEIRD" societies (Western, Educated, Industrialized, Rich and Democratic) (Henrich et al., 2010; see Ortega, 2005, and Andringa & Godfroid, 2020, for a review of this issue in the applied linguistics domain). Indeed, in many L2 speech sound learning experiments, the typical participant for an in-laboratory study is a college student from a WEIRD society, resulting in participants who are even less representative of the general population than if participants were sampled broadly from the same community. Further, it is likely that in-laboratory studies that are not part of a university course also select for a specific type of person who is, perhaps, especially interested in the research topic, organized enough to set up times to come into the lab, or may have other special personality features. However, these features may not be common among everyone who wants to (or needs to) learn additional languages.[1] Therefore, in-laboratory work has resulted in findings that are perhaps not robustly generalizable to a broader population.

In the present study, we treat the in-laboratory participants as the baseline and compare the remote group to them. This is in part because, as a field, we feel as though we have a better sense of what in-laboratory participants are like and what they can and cannot do as compared to remote participants. This is likely because we have decades of work with in-lab participants and relatively less experience with remote participants. However, it is possible that the in-lab participants are not a sufficient baseline for comparison. That is, in-person methods are likely

[1] Note that remote data collection may have similar challenges in terms of representation, given that individuals who seek out online studies may also not be representative of a broader population.

to create their own non-representative patterns or artifacts, especially because the laboratory context deviates significantly from a participants' daily experience. This has been demonstrated in previous work that has shown that being in a laboratory is sufficient to prime performance on a task using altered auditory feedback (e.g., Houde & Jordan, 2002).

1.2. Remote data collection in laboratory phonology

Because we are unaware of any work that directly compares remote data collection for learning novel speech sounds to in-laboratory learning, we address here the basic issues of remote data collection for laboratory phonology and psycholinguistic research more broadly, which underlie the type of training we conduct in the present study.

1.2.1. Pre-pandemic

Even before the COVID-19 pandemic necessitated the cessation of in-laboratory data collection and spurred many researchers to consider internet-based, remote data collection, researchers in linguistics, cognitive science, and psychology had already begun to make moves toward remote data collection to supplement or, in some cases, replace in-laboratory data collection. Many researchers championed the convenience of online data collection, especially in cases where in-person work was challenging. For example, Kimball and colleagues (2019) discussed ways to expand field studies and work with under-resourced languages using remote data collection. In other cases, recruitment of specific populations of participants within a community is challenging and thus remote data collection improves researchers' ability to investigate populations often not well-represented among college students (Staggs et al., 2022). It is also the case that some institutions and researchers do not have appropriate resources for in person data collection; therefore, remote data collection can democratize scientific investigation (Kimball, 2014).

However, even outside the recruitment of specific populations or the use of remote data collection to alleviate other challenging research environments, researchers in behavioral sciences have used online and remote data collection methods for many years. Many of these earlier studies focused on the feasibility of using online data collection tools, including Amazon's Mechanical Turk, Prolific, and FindingFive (Buhrmester et al., 2011; Crump et al., 2013; Mason & Suri, 2012), to conduct psycholinguistic research and replicate well-known in-laboratory effects. Indeed, many speech scientists began to use these tools to collect data and demonstrated that performance in these remote protocols was similar to in-laboratory experiments (e.g., Chodroff & Wilson, 2014; Yu & Lee, 2014). These tools were used for collecting perceptual judgments of speech (e.g., Kunath & Weinberger, 2010) and reaction time data (e.g., Enochson & Culbertson, 2015), among other measures.

The consensus before the pandemic was that these tools were extremely useful for data collection, as the typically onerous process of bringing participants into the lab to conduct studies could be made much easier by collecting data remotely. Researchers also noted other benefits, including recruiting participants from a more diverse population than what is typically available in a university human subject pool.

1.2.2. During the pandemic

Of course, at the height of the pandemic, interest in remote data collection expanded exponentially. Many (remote) conferences held special sessions or tutorials to demonstrate how to implement remote data collection (e.g., Rachel Theodore at the 2021 meeting of the Acoustical Society of America). With this proliferation of new studies, an interest in more closely comparing in person and remote data collection also grew. In an introduction to a special issue of *Linguistic Vanguard* around how remote data collection can be used effectively, Kostadinova and Gardner (2024) note that remote data collection often yields data on par with in-laboratory data collection. While many effects were replicated when data was collected remotely, some studies failed to replicate well-known results (Brekelmans et al., 2022; Grieve, 2021), leading to some skepticism from researchers and reviewers about the validity of online data collection. Further, remote vs. in-lab data collection interacts in important ways with task, suggesting that not all tasks are equally appropriate to be conducted online (Broś, 2025).

Questions about why such differences emerge has been a topic of substantial debate. For example, a recent paper demonstrated that participants tested in remote setups performed better on some tasks than participants completing tasks with an experimenter present (Bent et al., 2024). The authors suggest that these findings may be driven by the presence of an experimenter, which could shift participant behavior in a negative way or by demographic characteristics of the participants, which differed across the two data collection situations. Understanding whether and why differences might emerge across means of data collection is critically important if one hopes to use remote data collection to supplement or replace in-laboratory studies. Therefore, it is necessary to compare performance and logistical aspects of in-laboratory and remote data collection across a variety of tasks.

1.3. Current study

In this study, we compare in-laboratory and remotely collected data for training of discrimination of novel speech sounds. As described above, these types of training studies are often quite challenging to conduct in person because they typically require multiple days of training which have logistical and scheduling challenges for both participants and researchers. Therefore, if remote data collection yields similar results as in-laboratory data collection, one could feasibly replace or supplement data collection using logistically simpler procedures.

Below, we describe a two-day training study which was conducted both in-laboratory and using remote data collection methods. We present the results and discuss both methodological and theoretical implications of these results.

2. Methods

2.1. Participants

Participants for the in-laboratory condition of the experiment were recruited via the Linguistics and Psychology Human Subject Pool at the University of Oregon. We set a one-term window for recruitment, so all participants were recruited during a single 10-week quarter. Eighteen participants completed one day of the in-laboratory experiment and 16 of those participants completed both days of training.

Participants for the remote condition of the experiment were recruited via Prolific. We set a 30-participant cap for the experiment, and full completion of this condition of the study took two days. On day one, all participants were recruited and completed the first day of training, but not everyone took part in the training on day two. In total, 30 participants completed one day of training, and 22 participants completed both days of training.

All participants reported their ages as being between 18–35 years (mean = 21.6). They reported that their first language was English, and they were not proficient in any other languages. Further, no participants reported experience with any languages that use a dental-retroflex contrast (e.g., Hindi), which was used in this study (see Section 2.2, below). Participants in both conditions were paid for their participation.

2.2. Materials

Stimuli were drawn from a synthetic continuum originally created by Stevens and Blumstein (1975). They consist of a six-step continuum from / ɖa/ to / d̪a/, representing the Hindi dental-retroflex contrast. They vary in the onset of the second and third formant frequencies and in the frequency of the burst. Specific details about the creation of these stimuli can be found in Stevens and Blumstein. All stimuli were normed for duration and were amplitude normalized.

2.3. Procedure

The procedures for the in-laboratory and remote experiments were designed to be identical, except for location of the experiment. On the first day of training, participants began the experiment by giving informed consent for participation and completing a brief language background questionnaire to ensure they were members of our target demographic age range and language experience. Following this portion of the experiment, participants in the remote condition completed a headphone check (adapted from Woods et al., 2017) to ensure they were completing the experiment using headphones.[2] No participants were excluded based on the results of the headphone check.

Each of two days of training then followed the same pattern: Participants first completed a pre-test, then training, and then the post-test. During all portions of the experiment, participants completed an ABX discrimination task. They heard two different sounds from the continuum ("A" and "B") and then heard a third sound ("X"), which was identical to either "A" or "B". They were asked to identify which of the two sounds they heard by pressing either "f" or "j" on their keyboard.[3] The pre-and post-tests consisted of 72 trials each (a total of 288 trials across the pre- and post-tests on the two days). The training portion consisted of 504 trials per day (a total of 1008 trials across two days). This training and testing protocol is identical to those used in previous studies (e.g., Baese-Berk & Samuel, 2016, 2022).

2.4. Analysis

Data were analyzed following procedures in Baese-Berk and Samuel (2016, 2022) using logistic mixed effects regressions to estimate the proportion correct for discrimination of each pair of stimuli. The dependent variable is whether the participants responded correctly (coded as "1") or incorrectly (coded as "0") to the discrimination task.

The full model included a comparison of all the peripheral pairs to the center pair (i.e., coded so that the average of the other pairs was compared to Pair 3–4, with the non-central pairs as reference level), training modality (In-laboratory reference level vs. Remote), training phase (Day 1 Pre-Test reference level vs. Day 2 Post-Test), and interactions between these factors as well as a random intercept for participants.

Given previous work in categorical perception, the expectation was that participants should perform at chance for discrimination among all adjacent pairs at pre-test. At post-test, if participants have learned the novel speech sound contrast, they should perform at chance for pairs away from the category boundary (i.e., pairs 1–2, 2–3, 4–5, and 5–6) and above chance for the pair that crosses the category boundary (i.e., pair 3–4).[4]

[2] Interestingly some recent work suggests that use of headphones may not drastically impact some perception data (Sanker, 2023).

[3] The key correspondence to response was presented on the screen for each trial and has been previously used in other studies (e.g., Baese-Berk & Samuel, 2016).

[4] It should be noted that there has been substantial debate recently about the nature of categorical perception in general (e.g., McMurray, 2022) and in second language speech sound learning more specifically (Baese-Berk, Chandrasekaran, & Roark, 2022), including some work suggesting that learners are not actually acquiring categories. Further, the use of a single pair as the base pair has a long history; however, it is not without its own problems. We will return to these issues in the discussion section.

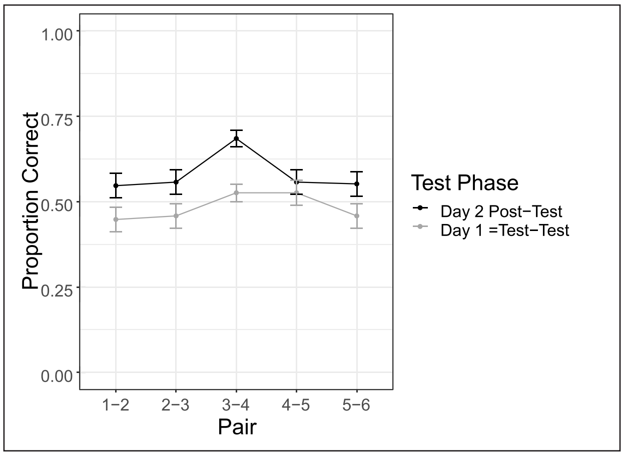

Figure 1: Average performance on Day 1 Pre-Test and Day 2 Post-Test for participants (n = 16) in the in-lab training condition. Error bars represent standard error of the mean.

3. Results

Below, we present the results of learning for each training condition separately, followed by a comparison of the two modalities. We begin with a general description of the trends in the data and follow this with the statistical analysis described above. Next, we present a discussion of individual performance and variability in this performance, and a discussion of the logistical aspects faced during completion of the experiments.

3.1. Learning after in-lab training

At pre-test, in-laboratory participants performed at chance for all pairs, as expected. Note the flat performance across all pairs in the grey line in **Figure 1**. This demonstrates that before training, participants were unable to discriminate between these unfamiliar speech sounds.

After training, we observe a "peak" for stimulus pair 3–4 in the black line in **Figure 1**, suggesting that in-laboratory participants have learned to differentiate between sounds that cross the trained category boundary, but not sounds within a single category. These results replicate decades of findings that after in-laboratory training, participants can learn to differentiate between two unfamiliar speech sounds.

3.2. Learning after remote training

When examining performance at pre-test for the remote participants, they perform at chance for all pairs, again as expected (see **Figure 2**). Again, this demonstrates that participants were unable to discriminate between these speech sounds before training.

At post-test, we again observe a "peak" at stimulus pair 3–4, showing that the remote participants learned to differentiate between sounds that cross the trained category boundary but not sounds within a single category (see **Figure 2**).

These results show that participants can learn novel speech sounds in remote experimental set ups. Below, we statistically compare learning across the two modalities.

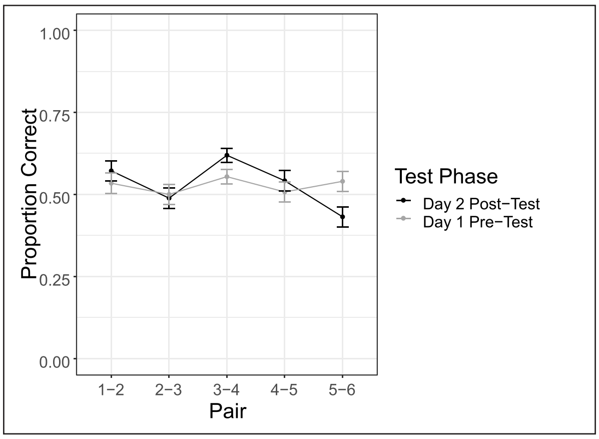

Figure 2: Average performance on Day 1 Pre-Test and Day 2 Post-Test for participants (n = 22) in the remote training condition. Error bars represent standard error of the mean.

3.3. Comparison of learning

Examining **Figures 1** and **2**, it is clear that participants in the in-laboratory condition show a higher discrimination peak than participants in the remote condition. Therefore, we used the logistic mixed effects model described above to investigate the role of training type on performance by predicting proportion correct as a function of test phase, pair, training type, and their interactions. The model is summarized in **Table 1** below.

Fixed Effect	**Estimate**	**Standard Error**	***z***	***p***
(Intercept)	0.234	0.078	3.025	0.002
Pair	0.368	0.092	4.003	<.001
Training Type	–0.169	0.101	–1.679	0.093
Test Phase	–0.184	0.073	–2.518	0.012
Pair × Training Type	–0.090	0.119	–0.761	0.448
Pair × Test Phase	–0.155	0.129	–1.203	0.229
Training Type × Test Phase	0.190	0.095	2.001	0.045
Pair × Training Type × Test Phase	–0.062	0.167	–0.373	0.70

Table 1: Model summary for logistic mixed effects model.

First, we see that pair is significant, demonstrating that participants discriminate between tokens 3 and 4 more accurately than other pairs along the continuum, which suggests categorical perception. Next, we see that the comparison between in-laboratory training and remote training is also significant, showing that participants in the in-laboratory training do, indeed, perform better than participants in the remote training group. Further, we see that test phase (i.e., Day 1 Pre-Test vs. Day 2 Post-Test) is significant, demonstrating a change in participant performance from Day 1 Pre-Test to Day 2 Post-Test. The interactions for pair by training type and pair by test phase are not significant; however, the interaction between training type and test phase is. This captures the observation above that participants in the in-laboratory training group are more accurate than participants in the remote training group at post-test, though they do not differ at pre-test. The three-way interaction is not significant.[5] It is possible that the lack of significant interactions is driven by individual differences among participants, and that investigating that variability will be informative. These statistical results suggest that participants do learn during training, and performance for in-laboratory participants is superior to remote participants. Below we explore in more detail why we may see these results.

3.3.1. Learning for individual participants

Previous work has demonstrated that individual performance among in-laboratory participants on this type of learning task differs significantly (e.g., Baese-Berk, 2019; Perrachione et al., 2011). In this section, we investigate individual performance on the post-test.

Specifically, we ask what proportion of participants demonstrate improvement from Day 1 Pre-Test to Day 2 Post-Test. For the purposes of this exploration, we define improvement from Day 1 to Day 2 as a change of .1 or greater on proportion of correct trials for the 3–4 pair. For the in-laboratory participants, 8 of 16 participants demonstrate a clear improvement from Day 1 to Day 2 for the 3–4 pair (see **Figure 3**). For the remote participants, only 9 of the 22 participants demonstrate a clear improvement (see **Figure 4**). In both cases, other participants demonstrate myriad other patterns, but do not demonstrate the canonical peak seen in the group data. A Fischer's exact test demonstrates that there is not a significant difference between the proportion of participants who learn (or demonstrate the canonical peak) across the two conditions ($p = .397$). This suggests that there is no evidence in this data that one type of training results in a higher fraction of individuals learning than the other.

[5] Please note that the conclusions one can draw from these simple effects are limited by the fact that simple effects apply at the default level of other variables. That is, the improvement from pre- to post-test holds for the in-laboratory group at pair 3–4 (i.e., the default levels for those two variables).

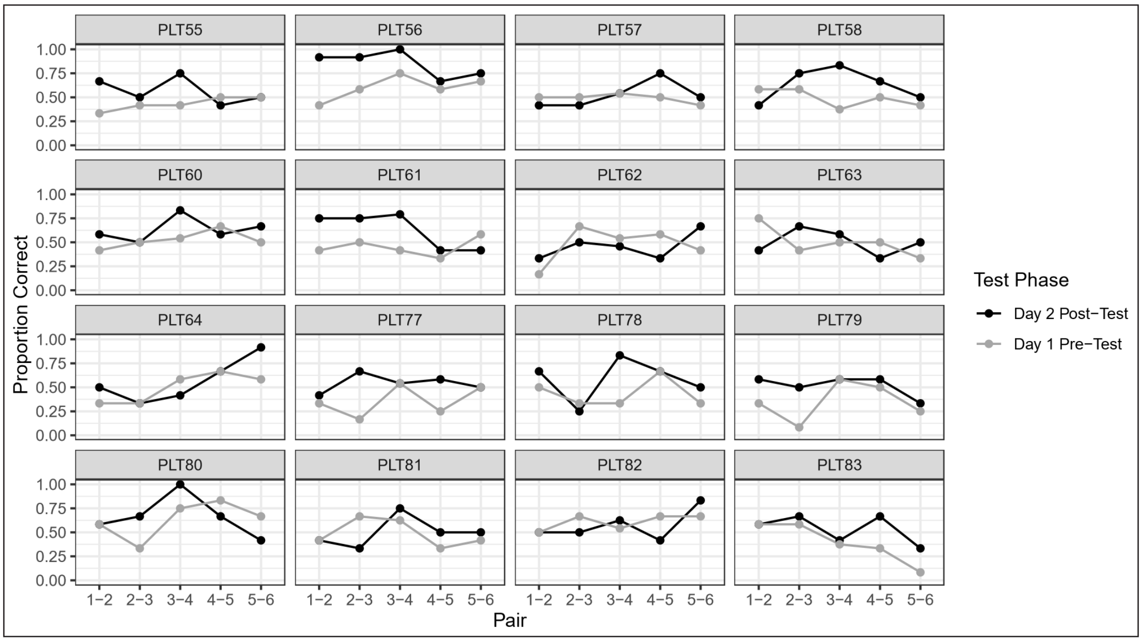

Figure 3: Individual performance on the Day 1 Pre-Test and Day 2 Post-Test for participants in the in-laboratory training condition.

3.3.2. Variability in performance

Next, we ask whether the variability in individual performance is different across the two groups. We again explored the difference in performance from Day 1 Pre-Test to Day 2 Post-test as a function of training condition. However, rather than asking a binary question of whether participants improve or not, we instead ask what patterns the group of participants show in terms of how much learning (or not) is demonstrated (see **Figure 5**).

Our initial hypothesis regarding variability was that the remote group would be more variable because they represent a more heterogeneous group of individuals. That is, participants who completed the experiment in the lab were all enrolled as college students at the University of Oregon and likely shared many demographic features. However, participants in the remote group were recruited from across the United States and likely varied more in their backgrounds, including in their current educational status. Interestingly, however, our data does not appear to be consistent with the original hypothesis. Using Levene's test for equality of variances, there is no difference between the two groups ($F = 4.0355$, $p = .052$). Examining **Figure 5**, we see that, if anything, the in-laboratory group demonstrates more variable performance than the remote group, not the predicted reverse pattern. Indeed, all participants in the remote group have counterparts in the in-lab group who demonstrate similar performance to them. We discuss the implications of this finding in more detail in the Discussion section, below.

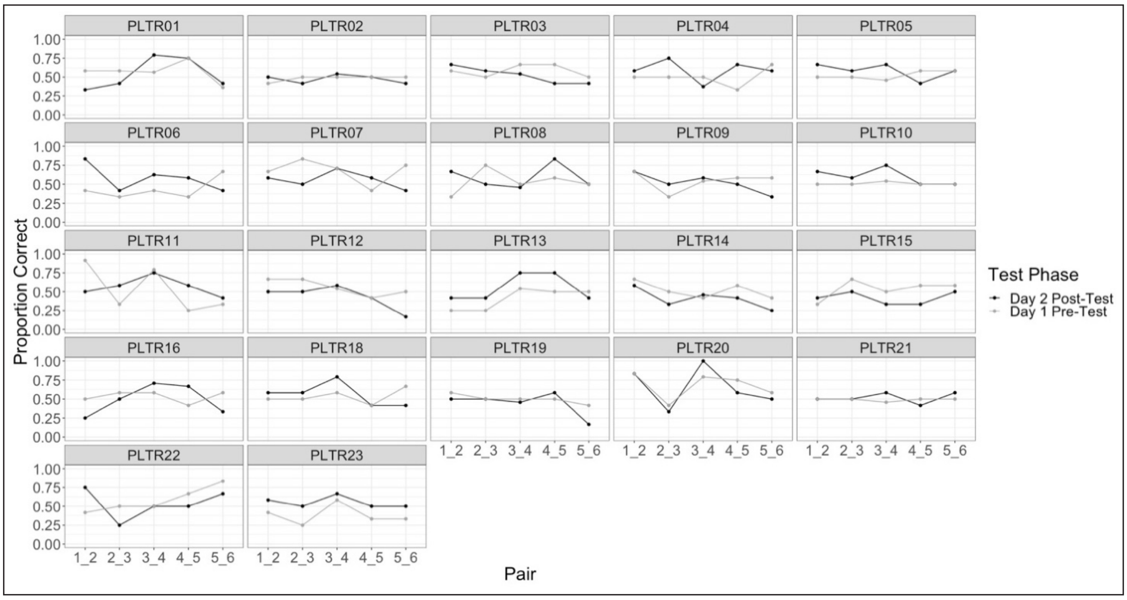

Figure 4: Individual performance on the Day 1 Pre-Test and Day 2 Post-Test for participants in the remote training condition.

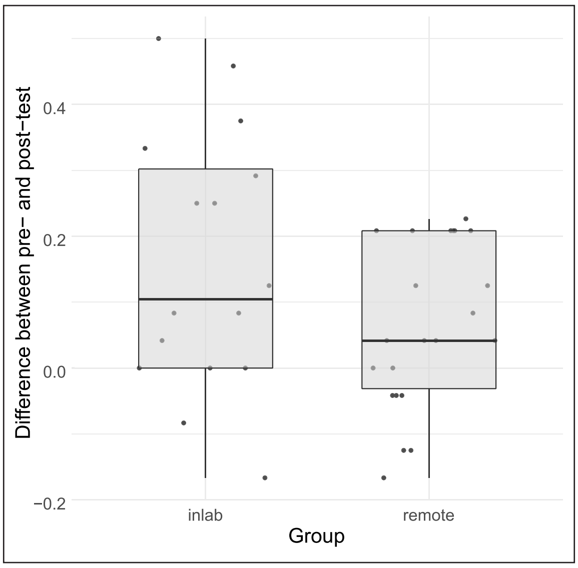

Figure 5: Difference between Day 1 Pre-Test and Day 2 Post-Test performance on the 3–4 pair for the in-laboratory and remote groups.

3.4. Comparison of logistical issues during training

When evaluating the two conditions for training, it is important to evaluate not only learning performance but also the logistics of conducting such experiments. One key consideration is the time it takes to recruit participants. In the case of the in-laboratory study, we recruited

18 participants in a single 10-week quarter. Sixteen of these participants completed the study. While myriad factors impact our ability to recruit participants, this recruitment pattern was quite typical for our lab for a two-day training study. It is often quite easy to recruit participants for shorter studies (i.e., single day studies), but for multiple day studies recruitment is typically much slower.

In contrast, participants in the remote condition were recruited on a single day. In fact, it only took three hours to recruit 30 participants, 22 of whom completed the study. Because the experiment took two days to complete, the data collection also took two days, but this time period was much shorter than the 10-week period described above. Indeed, if we had aimed for a much larger sample size, we could have recruited even more participants in the aforementioned three-hour time period.

It is also important to compare attrition rates across the two studies. In the in-laboratory condition, 2 of 18 participants did not return for the second day (around 11% of participants). In the remote condition, 8 of 30 participants did not complete the second day of training (27% of participants). While the attrition rate is higher for the remote condition, it is not remarkably higher, as attrition rates in our lab are often between 15 and 25% for multi-day training studies. Therefore, we do not believe that this is a significant disadvantage for the logistics of running such a study in a remote setting.

There were not significant differences between the groups in how long it took them to complete the task; all participants in both groups took around one hour each day to complete the task. Therefore, we do not believe that overall group differences in performance can be attributable to time-on-task.

One final question is one of demographics. In our data, our in-laboratory participants all identified as White (n = 16), and the majority of participants identified as female (n = 10; male n = 5; non-binary n = 1). This is not surprising given the demographics of the university and specifically of the Linguistics and Psychology Human Subject Pool we used to recruit participants.[6] The self-reported demographics of our remote participants were more ethnically diverse (White n = 18; Black n = 2; Asian n = 1; multiple racial ethnicities n = 1). The gender demographics were skewed more heavily toward male participants than our in-laboratory participants (male n = 13; female n = 7; prefer not to say = 2). Though we did not ask in our demographic questionnaire, it is also possible that our participant pools differed in other demographic characteristics including socioeconomic and educational status. The differences reported here between our in-lab and remote participant demographics are not substantial.

[6] Note that although the human subject pool used here included students from both linguistics and psychology classes, the students enrolled in this specific study were all currently enrolled in psychology classes and did not report experience in linguistic classes.

However, it is important to note that it is possible to actively recruit for more diverse populations in remote studies, which is more challenging for in-laboratory studies, especially in locations where the general population and, specifically, the university population of question, are more homogeneous.

4. Discussion

In this study, we examined learning of a novel phonological contrast after training conducted either in-laboratory or remotely. Our results demonstrate that, on a group level, both sets of participants demonstrated significant improvement from pre-test to post-test after training. However, participants in the remote condition demonstrated less learning than participants in the in-laboratory condition. Interestingly, participants in the remote condition were not more variable in their performance than participants in the in-laboratory condition, in spite of initial predictions to the contrary.

Our results demonstrate that training paradigms for novel speech sounds can be conducted remotely. As a group, our participants in both conditions improve from pre- to post-test. However, learning in the remote condition is less than that of the in-lab condition, which could result in some caution for researchers hoping to conduct such studies, especially given the wide individual variability among participants in both groups. However, in addition to some participants completing a remote training paradigm and others doing so in-laboratory, there are additional differences between the two populations here. This said, if comparison is within individuals participating in the same conditions (e.g., remote participants compared only to remote participants, rather than comparing in-lab participants to remote participants), it is likely that these differences do not preclude conducting experiments remotely, especially when the benefits of conducting experiments remotely are quite high (e.g., situations where data must be collected quickly or from populations not easily attainable for in-laboratory studies).

An additional factor to consider is that the population in the in-laboratory study is quite homogeneous. All participants were current students enrolled at the same university and who lived in the same geographic area. Therefore, education level was controlled and various other factors including race, socioeconomic status and personal background were unlikely to vary substantially given the demographics of both the region and institution. On the other hand, our remote participants were from across the United States, or at least were individuals using an IP address located within the United States. While they were roughly matched with the in-laboratory participants for age, all other factors were likely more variable for the remote participants. However, differences in demographics cannot fully account for differences in performance. For example, other studies that have matched in-laboratory and remote participant groups for a series of speech perception tasks have found that participants from the same population perform slightly differently in the two types of settings (e.g., Cooke & García Lecumberri, 2021).

As researchers who often crave control of our experimental population such that we do not introduce unnecessary variability in our data, this diversity may be a bit disconcerting. However, it is important to note that increasing the demographic variability in our population did not result in increased performance variability in the remote condition. Further, if the goal of psycholinguistic and laboratory phonology work is to capture generalizations about behavior or about language, we should question what it means when our results only hold for some subset of a population. That is, if our sample is not truly representative, what might this mean for our ability to make generalizations? Many recent papers have argued for more inclusive approaches to psychological and linguistic research and teaching (Baese-Berk & Reed, 2023; Higby et al., 2023; Kirk, 2023; Kutlu & Hayes-Harb, 2023; McMurray et al., 2023; Rad et al., 2018; Tripp & Munson, 2023). That said, increasing our samples beyond those which are most convenient (e.g., college students at our institutions) can feel daunting to researchers who are used to conducting in-laboratory studies with the convenient samples of individuals within their community. Using remote data collection protocols allows for relatively easy access to diverse populations. Indeed, in addition to collecting data from the general American public, remote data collection allows for research with a variety of specific populations. For example, collecting data from Spanish Heritage speakers in person was a very arduous process in our lab; however, when switching to remote data collection, we collected data from many such speakers, and these individuals had a substantially more heterogeneous background than those available to visit our laboratory in person (Staggs et al., 2022).

As discussed above, the logistics of conducting research remotely are often much easier than doing so in person. It can be challenging to schedule multi-day studies for participants at times that are also convenient or available for researchers. This is important because in addition to alleviating burdens on researchers in general, remote data collection may also allow for researchers who themselves are marginalized in a variety of ways to conduct research more easily. That is, individuals who are not at large institutions with human subject pools may find remote data collection allows them to complete studies relatively quickly that might otherwise take years to complete.

Even given all of the benefits of remote data collection described above, one could examine our data and still express concern that participants in the remote training condition performed less well than in-laboratory participants. That is, could this be a sign of less high-quality data in the remote group? For example, it is possible that participants in the remote condition are more distracted than participants in the in-laboratory condition. That is, participants in a laboratory setting tend to have a more controlled environment with fewer variables that may take a participants' attention away from the task at hand (see, e.g., Aivaz & Teodorescu, 2022). For example, smart phones have been shown to have a detrimental, distracting effect on learning in classrooms (Dontre, 2021), and the availability of such devices during remote learning could be responsible for the decrement in learning seen here. Alternatively, it could be the case that, in addition to demographic differences, remote participants are less accustomed to being in an educational or learning-oriented setting than our typical in-laboratory participants (e.g., Belot et al., 2015; Hooghe et al., 2010), and this lack of (recent) experience with educational settings and testing impacts performance.

We suggest that rather than viewing the decreased performance for remote participants compared to in-laboratory participants as a caution or a warning sign, we should view this as data—evidence that perhaps there are opportunities for interesting questions about how different groups of learners may perform differently, as well as what types of training might result in the most robust learning for different types of learners.

While the data presented here cannot differentiate between the effects of training modality and population differences, they do provide an opportunity to develop new research questions around how learners best acquire novel speech sounds. However, these questions are not just methodological. Theoretical questions can also be addressed by comparing both training conditions and participant populations. For example, significant previous work has suggested that individual variability in speech sound learning is not just unexplainable noise, but in fact correlates with various cognitive properties of the participant (e.g., working memory; McHaney et al., 2021; Roark et al., 2022; Perrachione et al., 2011). In some recent work, we have questioned whether all participants are, indeed, learning the same things during speech sound training tasks (Baese-Berk et al., 2022). That is, there is known to be substantial individual variation in performance both before, during, and after training on differentiating or categorizing novel speech sounds. Learners differ not only in performance but also strategies used (e.g., Chandrasekaran et al., 2014), weighting a variety of cues (e.g., Schertz et al., 2015), and various cognitive properties which may impact learning (e.g., Heffner & Myers, 2021). This leads to a question of whether participants in all cases are acquiring novel categories, and if not, what participants might be learning during training.

Indeed, this issue brings forth an even broader question about the nature of the assumptions we make in conducting studies like these. That is, a growing body of work questions the notion of categorical perception for speech (e.g., McMurray, 2022). Perhaps the variation we see in participant performance here could be informative about precisely what learners are acquiring and what our observations of how people learn can tell us about categorical structure, including interfaces of phonetics and phonology, broadly speaking. It is also possible that the use of a single pair (i.e., 3–4 in this study) is problematic because for some participants another pair crosses the category boundary instead. Indeed, if this is the case, one would expect less learning on a group level because good discrimination on one pair for one participant might be “canceled out” by poor performance by another participant. While this doesn’t appear to be the case in our data, it is possible that different participants have different category boundaries. This possibility is not often addressed in the literature because many studies assume (and have demonstrated) a natural psychophysical boundary for some contrasts (e.g., Elangovan & Stuart, 2008). While the effect of task on categorical perception has been well-studied (e.g., Kapnoula & McMurray, 2021; Pisoni & Tash, 1974; Schouten et al., 2003), the issue of addressing potential variance in individual performance as a function of different category boundaries, especially in cases of learning novel contrasts, is less understood. The consensus in the field seems to be shifting toward a need to better understand *what* precisely is being learned, and it is possible that using remote experiments will help address these issues.

4.1. Recommendations for conducting remote experiments

Taken together, the results of this study demonstrate the positive and negative aspects of conducting research remotely. While we are not the first to consider the trade-offs of these two means of data collection (see, e.g., Eerola et al., 2021), few studies have directly examined the pros and cons of conducting multi-day studies in these modalities and none has focused specifically on learning novel speech sounds. Below, we put forth a few recommendations for researchers deciding whether to conduct multi-day training studies remotely vs. in person.

First, researchers should consider logistical issues. If it is difficult to recruit participants to multi-day in-laboratory studies, or if they are hoping to target a specific population of participants (or a more diverse population than is available in their local area), remote data collection can provide a promising alternative to in-laboratory data collection.

Next, researchers should consider a number of methodological and analytical choices before beginning to conduct their research. In the present study, attrition was higher in the remote condition than the in-laboratory condition. How will the researcher handle attrition in the sample? Will they exclude participants who do not complete the entire experiment? Will they replace those who choose not to complete the study to ensure sufficient power? On the issue of power, the experimenter should consider whether a larger sample size is necessary to generate sufficient statistical power, as effect sizes may be smaller for remote experiments than for in-laboratory experiments.

The issue of potential distraction is also key to consider. How will a researcher determine whether a participant was too distracted? Or, more basically, how much does attention to task matter for the question being posed in a specific study? One could imagine a variety of controls for attention, including attention checks throughout the study, timed responses requiring participants to engage carefully on a given trial, or even use of remote eye-tracking tasks to note whether participants are looking at the screen during a given trial. The closer each task moves toward surveillance (e.g., a live experimenter being present via videoconference, for example), the more taxing this becomes for a laboratory and the more invasive the task becomes for participants (Castelli & Sarvary, 2021). This suggests a trade-off between controlled experimental settings and ease of remote data collection.

Further, the researchers should think carefully about how they collect data about their participants and whether verification of this data is mandatory in order to properly interpret their results. For example, it may be more difficult to verify some aspects of a participant's language background online than it is in person. If having a clear understanding of language background is mandatory for a specific experiment, researchers should consider whether additional measures may be required to ensure participants have a specific language background. However, it is also important to consider whether such information is truly necessary for a particular study, given that "nativeness," for example, is not a simple construct and may not impact study results as we expect (Cheng et al., 2021; Strand et al., 2024).

Given all of this, researchers should consider whether conducting data collection in a remote setting might be expected to impact their results in a meaningful way. That is, does shifting to remote data collection so drastically change the conditions for experimentation that it becomes a research question in and of itself?

Finally, and most crucially, researchers should be clear about their methodological and analytical decisions when reporting their results to ensure that results across studies are comparable and that results in any given paper are replicable.

5. Conclusion

In the present study, we demonstrate that participants in both in-laboratory and remote experimental settings can learn a novel speech sound distinction. However, differences in performance between the two groups suggest that remote and in-laboratory conditions are not identical and warrant significant consideration before replacing one with the other. While we believe that conducting laboratory phonology experiments remotely could improve representation of diversity in the populations we work with, it is not something that can be undertaken without significant consideration for how the change to remote settings may impact the results and interpretation of a given study.

Acknowledgements

This work is supported by NSF Grants BCS-2117665 and IIS-2024926. We would also like to thank Zachary Jaggers for his help in programming these experiments and Kurtis Foster for his assistance in data collection.

Competing interests

The authors have no competing interests to declare.

References

Aguinis, H., Villamor, I., & Ramani, R. (2021). Mturk research: Review and recommendations. *Journal of Management, 47*(4), 823–837.

Aivaz, K. A., & Teodorescu, D. (2022). College students' distractions from learning caused by multitasking in online vs. face-to-face classes: A case study at a public university in Romania. *International Journal of Environmental Research and Public Health*, *19*(18), 11188.

Andringa, S., & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics*, *40*, 134–142. https://doi.org/10.1017/S0267190520000033

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, *81*(4), 981–1005. https://doi.org/10.3758/s13414-019-01725-4

Baese-Berk, M. M., Chandrasekaran, B., & Roark, C. L. (2022). The nature of non-native speech sound representations. *Journal of the Acoustical Society of America*, *152*(5), 3025–3034. https://doi.org/10.1121/10.0015230

Baese-Berk, M. M., & Reed, P. E. (2023). Addressing diversity in speech science courses. *Journal of the Acoustical Society of America*, *154*(2), 918–925.

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23–36.

Baese-Berk, M. M., & Samuel, A. G. (2022). Just give it time: Differential effects of disruption and delay on perceptual learning. *Attention, Perception, & Psychophysics*, *84*(3), 960–980. https://doi.org/10.3758/s13414-022-02463-w

Belot, M., Duch, R., & Miller, L. (2015). A comprehensive comparison of students and non-students in classic experimental games. *Journal of Economic Behavior & Organization*, *113*, 26–33.

Bent, T., Lind-Combs, H., Holt, R. F., & Clopper, C. (2024). Perception of regional and nonnative accents: A comparison of museum laboratory and online data collection. *Linguistics Vanguard*, *9*(s4), 361–373. https://doi.org/10.1515/lingvan-2021-0157

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementaries. In O. S. Bohn (Ed.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/and /l/: Long-term retention of learning in perception and production. *Perception Psychophysics*, *61*(5), 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America*, *101*(4), 2299–2310. https://doi.org/10.1121/1.418276

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*, 104352.

Broś, K. (2025). Remote data collection in the study of ongoing sound change in Spanish – a comparative analysis. *Laboratory Phonology, 16*(1). https://doi.org/10.16995/labphon.10557

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Castelli, F. R., & Sarvary, M. A. (2021). Why students do not turn on their video cameras during online classes and an equitable and inclusive plan to encourage them to do so. *Ecology and Evolution*, *11*(8), 3565–3576.

Chandler, J., & Paolacci, G. (2017). Lie for a dime: When most prescreening responses are honest but most study participants are imposters. *Social Psychological and Personality Science, 8*(5), 500–508.

Chandrasekaran, B., Yi, H.-G., & Maddox, W. T. (2014). Dual-learning systems during speech category learning. *Psychonomic Bulletin & Review*, *21*(2), 488–495. https://doi.org/10.3758/s13423-013-0501-5

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

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *Journal of the Acoustical Society of America*, *136*(5), 2762–2772. https://doi.org/10.1121/1.4896470

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.

Crump, M. J., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PLOS One, 8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Darcy, I., Park, H., & Yang, C.-L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63–72. https://doi.org/10.1016/j.lindif.2015.04.005

Dontre, A. J. (2021). The influence of technology on academic distraction: A review. *Human Behavior and Emerging Technologies*, *3*(3), 379–390.

Earle, F. S., & Myers, E. B. (2015). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1680–1695. https://doi.org/10.1037/xhp0000113

Eerola, T., Armitage, J., Lavan, N., & Knight, S. (2021). Online data collection in auditory perception and cognition research: Recruitment, testing, data quality and ethical considerations. *Auditory Perception & Cognition*, *4*(3–4), 251–280. https://doi.org/10.1080/25742442.2021.2007718

Elangovan, S., & Stuart, A. (2008). Natural boundaries in gap detection are related to categorical perception of stop consonants. *Ear and Hearing*, *29*(5), 761–774.

Enochson, K., & Culbertson, J. (2015). Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PLOS One*, *10*(3), e0116946. https://doi.org/10.1371/journal.pone.0116946

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second language speech learning: Theoretical and empirical progress* (pp. 3–83). Cambridge University Press.

Grieve, J. (2021). Observation, experimentation, and replication in linguistics. *Linguistics*, *59*(5), 1343–1356.

Heffner, C. C., & Myers, E. B. (2021). Individual differences in phonetic plasticity across native and nonnative contexts. *Journal of Speech, Language, and Hearing Research*, *64*(10), 3720–3733. https://doi.org/10.1044/2021_JSLHR-21-00004

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

Higby, E., Gámez, E., & Mendoza, C. H. (2023). Challenging deficit frameworks in research on heritage language bilingualism. *Applied Psycholinguistics, 44*(4), 417–430. https://doi.org/10.1017/S0142716423000048.

Hooghe, M., Stolle, D., Mahéo, V. A., & Vissers, S. (2010). Why can't a student be more like an average person?: Sampling and attrition effects in social science field and laboratory experiments. *The Annals of the American Academy of Political and Social Science*, *628*(1), 85–96.

Houde, J., & Jordan, M. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research, 45*, 295–310.

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *Journal of the Acoustical Society of America*, *118*(5), 3267–3278. https://doi.org/10.1121/1.2062307

Kapnoula, E. C., & McMurray, B. (2021). Idiosyncratic use of bottom-up and top-down information leads to differences in speech perception flexibility: Converging evidence from ERPs and eye-tracking. *Brain and Language*, *223*, 105031.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2016). Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training. *Journal of Phonetics*, *57*, 21–39.

Kimball, A. E. (2014). The (statistical) power of Mechanical Turk. Purdue Linguistic Association Symposium. https://docs.lib.purdue.edu/plas/2014/proceedings/1/

Kimball, A. E., Keupdjio, H., Franich, K., & Kouankem, C. (2019). Expanding field studies using online speech perception experiments. In *Proceedings of the 19th International Congress of Phonetic Sciences* (pp. 315–319).

Kirk, N. W. (2023). MIND your language(s): Recognizing minority, indigenous, non-standard (ized), and dialect variety usage in "monolinguals." *Applied Psycholinguistics*, *44*(3), 358–364.

Kostadinova, V., & Gardner, M. H. (2024, January). Getting "good" data in a pandemic, part 1: Assessing the validity and quality of data collected remotely. *Linguistics Vanguard*, *9*(s4), 329–334. https://doi.org/10.1515/lingvan-2023-0170

Kunath, S., & Weinberger, S. H. (2010). The wisdom of the crowd's ear: Speech accent rating and annotation with Amazon Mechanical Turk. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 168–171). https://aclanthology.org/W10-0726.pdf

Kutlu, E., & Hayes-Harb, R. (2023). Towards a just and equitable applied psycholinguistics. *Applied Psycholinguistics*, *44*(3), 293–300.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*(3), 1242–1255.

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English/r/ and/l: A first report. *Journal of the Acoustical Society of America*, *89*(2), 874–886.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, *44*(1), 1–23. https://doi.org/10.3758/s13428-0110124-6

Matthews, G., & Campbell, S. E. (2009). Sustained performance under overload: Personality and individual differences in stress and coping. *Theoretical Issues in Ergonomics Science*, *10*(5), 417–442. https://doi.org/10.1080/14639220903106395

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*, 105010.

McMurray, B. (2022). The myth of categorical perception. *Journal of the Acoustical Society of America*, *152*(6), 3819–3842.

McMurray, B., Baxelbaum, K. S., Colby, S., & Tomblin, J. B. (2023). Understanding language processing in variable populations on their own terms: Towards a functionalist psycholinguistics of individual differences, development, and disorders. *Applied Psycholinguistics, 44*(4), 565–592.

Mora, J. C., & Darcy, I. (2023). Individual differences in attention control and the processing of phonological contrasts in a second language. *Phonetica*, *80*(3–4), 153–184. https://doi.org/10.1515/phon-2022-0020

Mora, J. C., & Mora-Plaza, I. (2019). Contributions of cognitive attention control to L2 speech learning. In A. M. Nyvad, M. Hejná, A. Højen, A. B. Jespersen, & M. H. Sørensen (Eds.), *A sound* *approach to language matters–In honor of Ocke-Schwen Bohn*, 477–499. Dept. of English, School of Communication & Culture, Aarhus University, Denmark. https://doi.org/10.7146/aul.322.218

Ortega, L. (2005). For what and for whom is our research? The ethical as transformative lens in instructed SLA. *Modern Language Journal*, *89*(3), 427–443. https://doi.org/10.1111/j.1540-4781.2005.00315.x

Pavlick, E., Post, M., Irvine, A., Kachaev, D., & Callison-Burch, C. (2014). The language demographics of Amazon Mechanical Turk. *Transactions of the Association for Computational Linguistics*, *2*, 79–92. https://doi.org/10.1162/tacl_a_00167

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America*, *130*(1), 461. https://doi.org/10.1121/1.3593366

Pisoni, D. B., & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception & Psychophysics*, *15*(2), 285–290.

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

Roark, C. L., Paulon, G., Rebaudo, G., McHaney, J. R., Sarkar, A., & Chandrasekaran, B. (2022). Individual differences in working memory impact the trajectory of non-native speech category learning. *PLOS One, 19*(6), e029717. https://doi.org/10.1371/journal.pone.2097917

Sanker, C. (2023). How do headphone checks impact perception data? *Laboratory Phonology*, *14*(1). https://doi.org/10.16995/labphon.8778

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. https://doi.org/10.1016/j.wocn.2015.07.003

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

Schouten, B., Gerrits, E., & Van Hessen, A. (2003). The end of categorical perception as we know it. *Speech Communication*, *41*(1), 71–80.

Staggs, C., Baese-Berk, M. M., & Nagle, C. (2022). The influence of social information on speech intelligibility within the Spanish Heritage community. *Languages*, *7*(3), 231. https://doi.org/10.3390/languages7030231

Stevens, K. N., & Blumstein, S. E. (1975). Quantal aspects of consonant production and perception: A study of retroflex stop consonants. *Journal of Phonetics*, *3*(4), 215–233. https://doi.org/10.1016/S0095-4470(19)31431-7

Strand, J. F., Brown, V. A., Sewell, K., Lin, Y., Lefkowitz, E., & Saksena, C. G. (2024). Assessing the effects of "native speaker" status on classic findings in speech research. *Journal of Experimental Psychology: General, 153*(12), 3027–3041. https://doi.org/10.1037/xge0001640

Strange, W., & Dittman, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics*, *36*, 131–145.

Tripp, A., & Munson, B. (2023). Acknowledging language variation and its power: Keys to justice and equity in applied psycholinguistics. *Applied Psycholinguistics*, *44*(4), 495–513. https://doi.org/10.1017/S0142716423000206

West, R. (1999). Visual distraction, working memory, and aging. *Memory & Cognition*, *27*(6), 1064–1072. https://doi.org/10.3758BF03201235

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

Yu, A. C., & 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. https://doi.org/10.1121/1.4883380