Laboratory Phonology
Journal of the Association for Laboratory Phonology

MacLeod, B. (2024). Variability and reliability in the AXB assessment of phonetic imitation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, 15(1), pp. 1–39. DOI: https://doi.org/10.16995/labphon.9379

Open Library of Humanities

Variability and reliability in the AXB assessment of phonetic imitation

Bethany MacLeod, School of Linguistics & Language Studies, Carleton University, Canada, beth.macleod@carleton.ca

Speakers adjust their pronunciation to come to sound more similar to recently heard speech in a phenomenon called phonetic imitation. The extent to which speakers imitate is commonly measured using the AXB perception task, which relies on the judgements of listeners. Despite its popularity, very few studies using the AXB assessment have considered variation or reliability in the listeners' performance. The current study applies a test-retest methodology focusing on the performance of listeners in the AXB assessment of imitation, which has not been considered explicitly before. Forty listeners completed the same AXB experiment twice, roughly two to four weeks apart. The findings showed that both sessions reach the same overall conclusion: The listeners perceived the same overall amount of imitation in both sessions, which is taken to mean that the shadowers did imitate and that the AXB task is reliable at the group level. Furthermore, the findings show that listeners vary substantially in their performance in the AXB assessment of imitation, but that they are relatively consistent in this performance across sessions. This suggests that differences in AXB performance at least partly reflect differences in ability to perceive imitation, rather than simply random variation.

1. Introduction

Speakers adjust their pronunciation to come to sound more similar to recently heard speech in a phenomenon called phonetic imitation.[1] Much of the existing research on phonetic imitation considers which acoustic-phonetic parameters talkers imitate, under which conditions, and which social, situational, cognitive, and linguistic factors mediate the process (e.g., Namy et al., 2002; Nielsen, 2011; Pardo, 2006; Yu et al., 2013; Zellou et al., 2017). The findings of these studies have been used to weigh in on a variety of central issues in linguistic inquiry, such as evidence for the nature of the link between perception and production, the nature of the phonetic representation of sounds, patterns of contact linguistics and sound change, and the acquisition of second languages (e.g., Lewandowski & Jilka, 2019; Pardo, 2012; Schertz & Johnson, 2022). In many studies, third-party judges (often called listeners or raters) listen to recordings of speech and decide if imitation has occurred (Babel et al., 2013; Babel et al., 2014; Clopper & Dossey, 2020; Dias & Rosenblum, 2016; Goldinger, 1998; Lewandowski & Nygaard, 2018; Miller et al., 2013; Namy et al., 2002; Nye & Fowler, 2003; Pardo et al., 2012; Pardo et al., 2018; Ross et al., 2021; Schertz & Johnson, 2022; Shockley et al., 2002; Walker & Campbell-Kibler, 2015). In this approach, the listeners' perceptual behaviour is treated as a tool or metric to measure the extent of imitation. Despite the popularity of the perceptual approach to assessing imitation, very few studies provide details about the behaviour of the listeners as individuals, instead usually only providing the group-level proportion of trials on which they detected imitation. Do individual listeners vary in their ability to perceive imitation? If they do, are they consistent in this ability? The answers to these questions matter for two reasons. The first reason is methodological. As noted above, the pattern of phonetic imitation is connected to several key debates in linguistics, and studies that use a perceptual assessment of imitation base their findings on the behaviour of the listeners. Given the importance of imitation studies for developing our understanding of these key issues, research that focuses on the listeners' behaviour is needed. We need to know more about the individual listeners who take part in these studies, how variable their ability to perceive imitation is, and how reliable findings are that use the perceptual assessment. The current study contributes to developing this knowledge by examining the reliability and consistency of individual variation in the perception of phonetic imitation in the AXB task.

The second reason we must explore individual variation in the perception of imitation relates to our understanding of why talkers imitate in the first place. Communication Accommodation Theory (CAT: Giles, 1973) proposes that the reason talkers imitate is social. When talkers imitate, this shrinks social distance between speakers, with the result that people who imitate are perceived as more likeable and attractive (e.g., Chartrand & Bargh, 1999). Furthermore, conversations in

[1] This phenomenon is also referred to as phonetic convergence, accommodation, alignment, entrainment, and synchrony. In this paper, I will use the terms phonetic imitation and convergence interchangeably.

which imitation takes place have been found to be rated more positively than those in which it does not occur (Giles & Smith, 1979; Street, 1982). Under CAT then, talkers imitate to influence their interlocutor's perception of them and their interaction. However, this influence can only take place if the interlocutor is able to perceive the imitation. As such, understanding more about how listeners perceive imitation and whether they differ in their ability to do so is a central piece in the puzzle of phonetic imitation. Solving this puzzle will help us generate insights not only about theoretical concerns such as the nature of the phonetic representation, but also about the social factors influencing the dynamics of interpersonal interactions (e.g., Giles et al., 1991).

2. Background

2.1. Measuring phonetic imitation

Speakers have been found to imitate a variety of acoustic-phonetic parameters such as vowel quality, vowel duration, word duration, speech rate, vowel nasalization, voice onset time, fundamental frequency (F0), and the acoustic realization of stress (Aubanel & Nguyen, 2020; Babel, 2010, 2012; Babel & Bulatov, 2011; Bonin et al., 2013; Brouwer et al., 2010; Clopper & Dossey, 2020; Cohen Priva et al., 2017; Cohen Priva & Sanker, 2018; Dufour & Nguyen, 2013; Kim & Clayards, 2019; MacLeod & Di Lonardo Burr, 2022; Nielsen, 2011; Pardo et al., 2017; Phillips & Clopper, 2011; Schweitzer & Walsh, 2016; Shockley et al., 2004; Walker & Campbell-Kibler, 2015; Zellou et al., 2017; Zellou et al., 2016). Phonetic imitation has been shown to occur both in conversation (Lewandowski & Jilka, 2019; MacLeod, 2014; Pardo, 2006) and in asocial, laboratory-based contexts (Babel et al., 2013; Goldinger, 1998; Kwon, 2021; MacLeod & Di Lonardo Burr, 2022; Nye & Fowler, 2003; Zellou et al., 2020). In lab-based studies, the shadowing task is a common way of eliciting imitation (Babel, 2010, 2012; Goldinger, 1998; Shockley et al., 2004; Walker & Campbell-Kibler, 2015). In this task, shadowers first provide a baseline of pronunciation, usually by reading aloud a word list, and then they shadow (repeat after) a model talker (a pre-recorded voice) saying the same words. The prediction is that the shadowers will shift their pronunciation when shadowing such that it comes to be more similar to the model talker than it was at baseline. Existing work has determined whether the shadowers have imitated the model talker using acoustic analysis of the recorded speech (e.g., Kwon, 2021) and/or a perceptual assessment, typically the AXB perception task (Babel et al., 2013; Babel et al., 2014; Clopper & Dossey, 2020; Dias & Rosenblum, 2016; Goldinger, 1998; Lewandowski & Nygaard, 2018; Miller et al., 2013; Namy et al., 2002; Nye & Fowler, 2003; Pardo et al., 2012; Pardo et al., 2018, Ross et al., 2021; Schertz & Johnson, 2022; Shockley et al., 2002; Walker & Campbell-Kibler, 2015, Zellou et al., 2020). On each trial in the AXB task, listeners hear triads of the same word, where X is the model talker and A and B are the baseline and shadowed productions from an earlier shadowing study, counterbalanced for order across trials. The listeners are asked to decide which of A or B sounds more like X; that is, which of the baseline or shadowed tokens sounds more like the model talker. If the overall percentage of trials on which the listeners choose the shadowed token is greater than 50%, we take this as evidence that the shadowers have imitated. The AXB approach is often heralded as superior to acoustic analysis due to its ability to better represent the perception of imitation in the real world and to allow listeners' perception to integrate across whichever acoustic parameters are present in the signal (Lewandowski & Nygaard, 2018; Miller et al., 2013; Pardo et al., 2012; Pardo et al., 2018).

2.2. The listener in the AXB task

The AXB task is used to assess perceptual discrimination patterns in a variety of areas of linguistic research. In most of them, the behaviour of the listeners is the object of inquiry. For example, the AXB task has been used to explore cross-language perception (e.g., Hallé et al., 2004; Polka, 1995) as well as the influence of factors such as age (Jia et al., 2006), regional accent (Larraza & Best, 2018), and age at which second language learning begins (Højen & Flege, 2006) on the perception of second language contrasts. In other cases, the listeners are not the object of inquiry but rather serve as a tool to assess a language pattern in the speakers they listen to. For example, in studies in the speech and language pathology literature, listeners evaluate recorded voices for characteristics such as pathological voice quality, typically with a real-world application such as detecting voice disorders (e.g., de Krom, 1994). Similarly, in studies of phonetic imitation, the behaviour of listeners in the AXB task is used to explore the extent to which speakers (such as shadowers from a shadowing experiment) have imitated (e.g., Dias et al., 2021). While the listeners are the foundation of the AXB task, most existing work has provided few details about their behaviour beyond the overall percentage (calculated at the group level) of AXB trials on which they detected imitation (that is, chose the shadowed token as being more similar to the model). With that percentage reported, most studies move directly to statistical modeling to determine if the overall percentage is statistically significantly higher than 50% and/or if it is influenced by a variety of other factors, such as lexical frequency of the stimuli (e.g., Dias & Rosenblum, 2016; Goldinger, 1998) or dialect or gender of speaker or model (e.g., Babel et al., 2014; Namy et al., 2002; Pardo, 2006; Pardo et al., 2017; Ross et al., 2021; Walker & Campbell-Kibler, 2015). We are rarely provided with any information about how the individual listeners might have varied in their performance in the AXB task.

Despite this, a small number of studies offer intriguing hints of variation in listeners' abilities to perceive imitation in the AXB task. For example, Babel & Bulatov (2011) examine the imitation of fundamental frequency (F0) and assess imitation using both acoustic analysis of F0 and an AXB perception task with a group of 20 listeners. In addition to results about the production of imitation by a group of shadowers and the influence of shadower gender and condition (high-pass filtered versus unfiltered), their findings also suggest that the AXB listeners "demonstrate different propensities to detect similarity in an AXB task" (Babel & Bulatov, 2011,

p. 13), with individuals ranging from choosing the shadowed token in roughly 45% of trials to roughly 55% of trials. They conclude that findings from perceptual tests, such as the AXB task, will be influenced by individual differences in perceptual sensitivity such that listeners will differ in their ability to perceive imitation. More recently, Schertz & Johnson (2022) found variation among listeners in a variant of the AXB task, the ABX task. In that study, the ABX task was used to evaluate participants' perception of variation in VOT, which was then compared to the extent to which they themselves would imitate that variation in VOT. Although the individual perceptual pattern is not discussed explicitly, the participants from Schertz & Johnson (2022) still exhibited a substantial amount of individual variation (as shown in their Figure 4), with individual participant means varying from approximately 15% correct to 100% correct.

It is perhaps not surprising that listeners would vary in their ability to perceive imitation. Individual variation in perception has been observed in a variety of other lines of research including use of cue weights in perception of first language contrasts (e.g., Kong & Edwards, 2016; Schertz et al., 2015), perception of second language contrasts (e.g., Kim et al., 2017; Mayr & Escudero, 2010), and perception of speaker-specific phonetic detail (see Smith, 2015, for a review). In fact, Pardo (2013) explains that just as speakers might be expected to imitate to different extents, listeners would likely also vary in their evaluation of similarity in such a perceptual task. Pardo suggests that it is therefore important to include many listeners in an AXB task and to "incorporate all levels of variability in the analysis" (Pardo, 2013, p. 2). This seems to mean that when analyzing the pattern of the perception of imitation from the AXB task, we should include factors that can capture individual listener variation. In fact, many studies have done this, typically by including a random intercept for listener in a logistic mixed effects model where the binary choice of baseline or shadowed token is the dependent variable. However, even when this random effect is included, most studies do not discuss the impact of including that effect on the model (Babel et al., 2013; Babel et al., 2014; Clopper & Dossey, 2020; Dias & Rosenblum, 2016; Lewandowski & Nygaard, 2018; Miller et al., 2013; Pardo et al., 2018). Without this discussion, it is difficult to know to what extent there is variation among the listeners in their ability to perceive imitation.

2.3. Individual differences and reliability

Given that listeners have been found to vary in perceptual processing in many ways (see above), we might predict that listeners will also vary in their ability to perceive imitation in an AXB task. Furthering our understanding of the pattern of individual differences in phonological processing contributes to discussions of key issues such as the nature of the phonetics-phonology interface, first and second language acquisition, and patterns of sound change (Wade et al., 2020; Yu & Zellou, 2019). Of course, some variation is expected among individuals simply due to the reality of dealing with human behaviour. To determine whether this variation represents a stable property of the individuals (i.e., genuine individual variation) and not random variation, we must demonstrate that the individual variation is consistent (Cohen Priva & Sanker, 2020; Kong & Edwards, 2016; Wade et al., 2020). To do this, we test the same participants on the same task on two occasions and then consider the extent of individual variation in each session and how that variation relates across sessions. If the individual differences in each session were random, we would not expect a relationship in individual patterns across sessions. This approach provides a measure of the consistency or *test-retest reliability* of the participants' performance in the task.

Test-retest reliability is one of several types of reliability that can be assessed, along with interrater reliability, parallel forms reliability, and internal consistency (e.g., Bannigan & Watson, 2009). Test-retest reliability is a measure of how consistent a test is over time. As noted above, to evaluate this kind of reliability, researchers carry out the same task with the same participants at two time points and determine the relationship between the participants' performance. Once the two sessions have been carried out, the test-retest reliability is quantified using the Pearson correlation between the scores on the two tests (Bannigan & Watson, 2009; Carmines & Zeller, 1979, p.38; DeVon et al., 2007): the higher the correlation, the higher the test-retest reliability. This is usually thought of as a characteristic of the test (DeVon et al., 2007) and not so much a characteristic of the individuals who take the test. In this way, we can think of the test-retest reliability as a way of evaluating the reliability of the AXB task as an instrument for assessing whether a group of shadowers has imitated a model talker. The test-retest reliability would be high if the conclusion we reach about whether the shadowers imitated or not (as a group) is the same in both sessions. In addition to this approach, however, we can also think of test-retest reliability as a characteristic of the individuals who take the test. Kong & Edwards (2016) used a test-retest approach to explore individual variation in categorical perception and the use of VOT and F0 as cues to perceiving the stop voicing contrast in English. They found that the extent to which listeners perceived the contrast as gradient or categorical and their sensitivity to F0 was consistent across sessions. Within imitation, Wade et al. (2020) used a test-retest design to explore the stability of individual differences among shadowers in the imitation of extended VOT. Their results showed that the degree of imitation by individuals was highly correlated across sessions, suggesting that imitative tendency (at least for extended VOT in a shadowing task) reflects a stable characteristic of individual talkers and not simply random variation. The current study applies a similar test-retest methodology but focuses on the performance of listeners in the AXB task to assess imitation, which has not been considered explicitly before.

2.4. The current study

The current study has three goals, listed below.

1. Reliability of the AXB task: to determine whether the AXB task is a reliable method of measuring phonetic imitation using approaches commonly taken in the literature,

2. Individual variation in ability to perceive imitation: to explore the degree of individual variability in listener performance in the AXB task,
3. Consistency of individual ability: to determine how consistent the listeners' performance is across sessions.

To achieve these goals, this study applies the typical methodology of using an AXB perception task to assess the amount of imitation produced by shadowers from a shadowing task; however, unlike previous work, the AXB task was repeated in a second session with the same group of listeners, 12 to 27 days later. The findings showed that both sessions reach the same overall conclusion: the listeners perceived the same overall amount of imitation in both sessions, which is taken to mean that the shadowers did imitate and that the AXB task is reliable at the group level. Furthermore, the findings show that listeners vary substantially in their performance in the AXB assessment of imitation, but that they are relatively consistent in this performance across sessions. This suggests that differences in AXB performance at least partly reflect differences in ability to perceive imitation, rather than simply random variation.

3. Methodology

This section discusses the creation of the AXB materials and the running of the experiment. As noted above, the goal was to apply a methodology that is consistent with much of the existing literature on phonetic imitation. To that end, a shadowing study was run using the typical approach and then the recordings from that study were used as the materials in the subsequent AXB perception experiment.

3.1. Creating the AXB materials

All recordings (model talkers and shadowers) were made with participants seated in a sound-attenuated booth in front of a small desk that held a computer monitor. The recordings were made using a Sound Devices Mix-Pre 6 II recorder and Audio Technica AT831b lavaliere microphone, which was attached to a small mic stand placed on the desk beside the computer monitor. The stimuli were presented using OpenSesame (Mathôt et al., 2012). All participants were compensated with $10 CAD per session.

3.1.1. Stimuli

The stimuli in the word list (provided in **Table 1**) were a set of 20 disyllabic English words, all with stress on the first syllable, half low frequency and half high, determined using the Zipf metric (van Heuven et al., 2014) in the SUBTLEX-US frequency list (Brysbaert & New, 2009). Note, however, that while previous work has suggested that lexical frequency could influence the degree of imitation (e.g., Babel, 2010; Dias & Rosenblum; Pardo et al., 2017, but see Pardo et al., 2013), it is not the purpose of this study to explore the effect of lexical frequency on the production or perception of imitation. Instead, the stimuli are controlled for frequency to potentially generate instances of different amounts of imitation to which the listeners in the AXB tasks could respond.

Low frequency	High frequency
toucan	raising
pinkish	harder
cyclist	table
dwindle	music
snippet	welcome
sprinter	moment
foodie	question
glacial	morning
larder	people
boatyard	never

Table 1: List of stimuli.

3.1.2. Recording the model talker

To create the model talker recordings, four talkers were recorded reading the word list aloud. The stimuli appeared on the computer screen one at a time, changing to the next word automatically every three seconds. The talkers read the word list aloud five times, in random order each time and were recorded. All were female, monolingual speakers of Canadian English, aged 23 to 24, who grew up in Sault Ste-Marie, ON,[2] which is a city of approximately 74,000 people in Northern Ontario just across the border from Michigan on the St. Mary's River, near the southeast end of Lake Superior. Model talkers from this region were included to introduce some subtle dialectal variation that could potentially be imitated by Ottawa-local shadowers who do not speak a Northern variety. Two such variations are captured by the word list in this study: 1. the phonetic diphthongs /eɪ/ and /oʊ/ were realized as monophthongs in words such as *glacial* [gle.ʃəl] and *boatyard* [bot.jæɹd]; 2. the low, back vowel was fronted before /ɹ/ in words such as *harder* [hæɹ.dɚ]. Having these differences is desirable since it allows the shadowers the opportunity to perceive the variation and potentially imitate it during the shadowing task. This improves the chances that the listeners will have some imitation to perceive. Most phonetic imitation studies introduce some kind of variation between model talker and shadower stemming from differences

[2] The talkers were all living in the local (Ottawa, ON) area at the time, making recording their voices in person straightforward.

such as dialect (e.g., MacLeod, 2014; Walker & Campbell-Kibler, 2015), gender (e.g., Babel & Bulatov, 2011) or age (e.g., Lin et al., 2021), although other studies include speakers of the same dialect (e.g., MacLeod & Di Lonardo Burr, 2022) and explore imitation of more idiosyncratic characteristics. Similar to the note above about lexical frequency, while the model talker and shadowers have some minor dialectal differences in their pronunciation, this study does not aim to explore the effects of that variation on either the production or perception of imitation. Instead, that variation is included only to generate targets for imitation beyond idiosyncratic differences.

From the four possible model talkers, one was chosen to serve as the model talker in subsequent phases. This particular talker was chosen because her recordings were free of extraneous noises, such as coughing, and she produced the words in a clear voice without whispering or using creaky voice. The model talker recordings to be used in the shadowing task and subsequent AXB task were taken from the fourth and fifth repetitions of the word list, choosing individual recordings that were clearest and best captured the dialectal variants described above (Ross et al., 2021, p. 6). Individual sound files were normalized to an intensity of 70dB.

3.1.3. Recording the shadowers

Ten talkers took part in the shadowing phase. All were female, first-language speakers of Canadian English[3], ranging in age from 18 to 21, with a median age of 18. Of these 10, two were excluded, one for having no hearing in one ear and the other for being from Sault Ste-Marie, ON. The remaining eight shadowers were from different parts of Ontario or Quebec, but none was from Northern Ontario. Previous work has considered the effect of gender on phonetic imitation and, while the results are somewhat contradictory (see Pardo et al., 2017 for a review), it seems there could be an effect. This study is focused on the perception of imitation, so to avoid influences of gender, only female model talkers and shadowers were included.

There were three phases in the shadowing experiment. The first phase is the baseline, where shadowers read aloud the list of stimuli to provide an estimation of their pre-exposure pronunciation. They read the list of words five times, in random order each time. The second phase is the shadowing, where the shadowers heard recordings of the model talker producing the stimuli and they repeated the words immediately after each ended. The model talker recordings were presented to shadowers over speakers instead of headphones to allow the participants to monitor their own voice without hearing it through the physical barrier of the headphones.[4]

[3] In addition to English, some also had other first languages, including French, Arabic, and Hindi.

[4] This approach also allows us to measure the duration of the interval between the offset of the model production and the onset of the shadowing. The duration of this interval could be related to how much the shadowers imitate, although this factor was found not to influence the perception of imitation in MacLeod & Di Lonardo Burr (2022). This idea is not explored in the current paper, but the information could be useful in future studies that use this dataset.

Shadowers were instructed to say each word they heard aloud. They shadowed all 20 words, three times, in random order each time. The third and final phase is the post-exposure phase, where the shadowers produced the word list once more. Only the baseline and shadowing phases are needed in the current paper.

The recordings of four of the eight shadowers were chosen to be included as stimuli in the AXB task. These particular four shadowers were chosen because they were deemed to include varying degrees of phonetic imitation, as determined perceptually by the author and two research assistants. Having this variation allows an investigation into whether listeners perceive different amounts of imitation among the shadowers and how consistently listeners order the shadowers with respect to how much they imitated. The recordings of the second baseline reading were used as the baseline stimuli in the AXB task to avoid any issues with hyperarticulation in the first round, while the third round of shadowing were used since previous work has suggested that more exposures to a pronunciation can increase the amount of imitation (Goldinger, 1998; Miller et al., 2013, cf. Babel et al., 2013 and see Lewandowski & Nygaard, 2018: 624). Individual sound files were normalized to an intensity of 70dB before being used in the AXB experiment.

3.2. AXB experiment

3.2.1. Participants

Forty participants[5] (34 female, four male, three non-binary) took part in the AXB perception task as listeners.[6] All were first-language speakers of Canadian English (19 to 68 years of age[7]: mean 26, median 22) with no reported hearing or speech disability.

3.2.2. Procedure

Due to the COVID-19 pandemic, the AXB perception task was run online via Zoom. Once OpenSesame was launched, the researcher and the listener joined a Zoom call and the researcher used the remote control function to give the listener remote access to the OpenSesame window. This allowed the listeners to enter their responses directly onto the researcher's computer.[8]

[5] A total of 41 participants took part, but one was removed due to failing to complete the second session.

[6] In the interest of normalizing providing such information, the listeners reported the following ethnicities: Asian and White (1), Arab (3), Hispanic (2), South Asian (2), White (32).

[7] While the age range is quite wide, 90% of the participants (36/40) were born in 1988 or later.

[8] While this approach allowed the study to be carried out in a safe way, it introduces a certain amount of variation in the procedure since participants each use their own headphones and are situated in different environments. These differences could contribute to individual variation in performance in the AXB task. I was careful to ensure that the participants always used headphones (as opposed to speakers), and used the same headphones and were in the same location in both sessions. In this way, any differences between participants stemming from differences in equipment or environment was held constant between the two sessions, minimizing any impact on the participants' consistency in the AXB task.

In the AXB task to assess phonetic imitation, participants hear three repetitions of a word on each trial: A X B. The model talker production is in the X position and A and B are a shadower's production of the same word in either the baseline or shadowing phase. The order of the shadower productions is counterbalanced across trials. The participants' task is to listen to the pronunciation of the words in each trial and decide whether the first word (A) or the last word (B) sounds more similar to the middle word (X). In the current study, participants were asked to focus on pronunciation as opposed to other things, such as background noise (Miller et al., 2013; Pardo, 2006), which was minimal to begin with. Participants first completed 10 practice trials, followed by a pause in which they could ask any questions, and then completed a further 5 practice trials. This two-phase practice was included to ensure that participants understood the purpose of the task and how to complete it and allowed the researcher to verify that the sound was at an appropriate volume by asking the listener. All practice trials involved different shadowers from those included in the test trials. Next, the listeners completed the test trials. The experiment proceeded to the next trial after a response was logged or after 7 seconds had elapsed with no response. All 20 words from each of the four shadowers were included twice, once with the shadowed production in the A position and once in the B position. Each trial was also included twice (two exposures) for a total of 320 trials (20 words x 4 shadowers x 2 AXB orders x 2 exposures), which were split into four blocks of 80 trials. Participants were offered a break between blocks.

All listeners heard the trials in the same order, which was not completely random. First, in terms of the shadowers, the order was set to cycle through each shadower. That is, the first trial was Shadower A, followed by Shadower B, followed by Shadower C and then Shadower D, and then back to Shadower A. Pilot testing of the AXB task suggested that when only one shadower was presented in a block, listeners could come to notice specific aspects of the baseline versus shadowed pronunciation and develop a strategy to respond to the trials without attending to the model talker pronunciation. For example, if a shadower produces the baseline trials with a rising intonation, but imitates a model talker's falling intonation, listeners could come to notice this and then apply that knowledge in later trials without listening to the model talker. That is, they would always choose the token with falling intonation. Although such a strategy would likely result in the listener choosing the shadowed token, it would not reflect the purpose of the AXB task, which is to judge similarity on each trial. To avoid listeners developing such strategies, all four shadowers were included in each block and each trial presented a different shadower from the one before it. The words themselves were pseudorandomized, alternating between high and low frequency words and avoiding adjacent similar sounding words, such as *harder* and *larder*. The position of the shadowed token in the triads (i.e., in the A or B position) was randomized across trials. The interstimulus interval was 150ms, which should allow listeners to make use of fine-grained phonetic information in making their judgements (Werker & Logan, 1985). All

40 listeners returned to complete a second session of the AXB experiment 12 to 27 days later (median: 16 days). The task was identical to the one completed in the first session, including trial order.

3.2.3. Data analysis

In total, 25,600 AXB trials were collected (320 trials x 2 sessions x 40 participants). Trials that met the following criteria were removed: those with no response (51, 0.2% of the total), those with response times more than two standard deviations from the mean, calculated individually by listener (1,125, 4.4% of the data), and those with response times under 150ms (88, 0.3%). This left 24,336 trials across both sessions. To explore listener variability and consistency, a response from both sessions is needed. Any trials that were missing a response from one session (typically due to a long response time) could not be used, even if the response from the other session was available. In the end, there were 11,555 pairs of responses (23,110 total trials) from sessions 1 and 2.

To stay consistent with the statistical approach used in most work on phonetic imitation, frequentist logistic mixed effect modelling was employed, using the *lme4* package (Bates et al., 2015). The *lmerTest* package (Kuznetsova et al., 2017) was used to calculate p-values for the logistic models, using degrees of freedom based on the Satterthwaite approximation. Models included random intercepts for listener, shadower, and word and by-listener slopes for any significant fixed effects (Sonderegger, Wagner & Torreira, 2018). Plots were generated using the package *ggplot2* (Wickham, 2016).

4. Results

This section discusses the results of the experiment, focusing on two main areas. Section 4.1 considers the reliability of the overall pattern of the perception of imitation as a way of establishing the reliability of the AXB task for assessing imitation. Section 4.2 focuses on individual variation in the perception of imitation in order to explore to what extent individuals vary in their ability to perceive imitation and their consistency in doing so across sessions.

4.1. Reliability of AXB task for assessing imitation

In phonetic imitation studies, the results of an AXB task are usually used to determine if the shadowers have imitated (rather than focusing on the perception of imitation itself). In studies of phonetic imitation that use the AXB assessment, the results are often presented with the following two elements: (1) The overall percentage of trials in which the listeners chose the shadowed token as being more similar to the model talker and (2) A statistical model that determines if that overall percentage is statistically significantly higher than 50% by examining the sign and significance of the intercept (as in, for example, Pardo et al., 2017). The percentage of trials on which the listeners chose the shadowed token is often referred to as the percentage of correct responses (%-correct). Of course, in an AXB task using recordings from a shadowing task, there really are no correct or incorrect answers; instead, listeners decide whether the baseline token or the shadowed token sounds more similar to the model talker and either response is plausible since shadowers can diverge as well as converge (Babel, 2010; Giles, 1973). However, using the term %-correct is a convenient and common shorthand that I will make use of in this paper.

Here, I am interested in knowing if the conclusion about whether the shadowers imitated would be the same in both sessions. That is, are the findings of the AXB task reliable across sessions? The overall %-correct across both sessions was 58.8% and the proportions are extremely similar in the two sessions, with 58.7% in Session 1 and 58.9% in Session 2. These percentages are near the upper end of group-level values typically reported in perceptual assessments of imitation, which generally range from 52% to 58% (Kim, 2012; MacLeod & Di Lonardo Burr, 2022; Nielsen & Scarborough, 2019; Pardo et al., 2017; Shockley et al., 2004; Wagner et al., 2021; Walker & Campbell-Kibler, 2015). To determine whether the overall group-level proportion of 58.8% reflects evidence that the listeners perceived imitation, a multilevel logistic regression model was fitted to the AXB responses using the glmer() function in the *lme4* package (Bates et al., 2015). I included three fixed effects: SESSION, to determine if the pattern of perception differed between the sessions; AXB ORDER, to explore whether the listeners' responses depended on the position of the shadowed token in the A or B position; and EXPOSURE, to capture the fact that each trial was repeated twice in the perception experiment. The listeners' responses were dummy coded as a binary dependent variable: 1 on trials where the listener selected the shadowed token and 0 where the listener selected the baseline token. The fixed effects were sum coded, allowing an interpretation of the intercept with all factors as their average. The model, shown in (1), included random intercepts for *listener, shadower,* and *word* and a by-listener slope for AXB ORDER.

(1) glmer (CorrectYN ~ 1 + SESSION + AXB ORDER + EXPOSURE
+ (1+ AXB ORDER|listener) + (1|shadower) + (1|word)

The results are provided in **Table 2**. The intercept is positive and significant. This indicates that the proportion of trials in which the listeners chose the shadowed token is significantly higher than 50%. As is standard practice in studies that use the AXB method of assessing imitation, we can conclude that the shadowers did imitate the model talker at the group level, taking both sessions together. Furthermore, the effect of session was not significant. This confirms our impression that the overall %-correct is essentially the same in both sessions (58.7% in Session 1 and 58.9% in Session 2). The main effect of AXB ORDER was not significant, but a comparison of the model in (1) to one without each of the by-listener slope for AXB ORDER indicated that the model in (1) provided a better fit. Examining the by-listener coefficients for AXB ORDER revealed that they ranged from –2.12 to 1.96. Sixteen of the 40 listeners had negative slopes, indicating that they were more likely to choose the shadowed token when it was in the B position (BS order). The remaining 24 listeners had positive slopes, meaning they were more likely to choose the shadowed token when it was in the A position (SB order). Others have suggested that AXB ORDER influences listener responses (e.g., Pardo et al., 2013; MacLeod & Di Lonardo Burr, 2022), but the finding here points towards this effect being more individual rather than universal.

	β	*SE*	*p*
Intercept	0.393	0.17	0.024*
SESSION	0.009	0.03	0.747
AXB ORDER	0.144	0.14	0.316
EXPOSURE	–0.031	0.03	0.263

Table 2: Statistical results from logistic mixed effects model for AXB results across both sessions.

Taken together, these results indicate that the main conclusion of overall imitation does not depend on session. This suggests that, at least for this purpose, the AXB task is a reliable tool for assessing group-level imitation.

4.1.1. Replication of experiment using bootstrapping

A reviewer suggests another way of testing the reliability of the AXB task, which is to replicate the experiment with more than one group of participants and compare the findings across groups. We can test this idea without rerunning the experiment using the bootstrapping procedure, in which samples are taken from the original dataset, tested in some way, and the results compared. This procedure was applied in the current dataset in the following way. First, I randomly sampled 20 of the 40 listeners, creating Sample 1, then put the remaining 20 listeners into Sample 2. Next, I applied the statistical model in (1) used to determine if the listeners perceived imitation to each sample. Lastly, I compared the significance of the intercept for each sample. If this procedure is repeated many times, we can determine the proportion of times in which the outcome of the models match across the two samples. If this proportion is high, we can take this as evidence of the reliability of the AXB task.

The procedure above was implemented in R and was repeated 1000 times. Of the 1000 runs, 978 of the sample pairs reached the same conclusion, which was that the listeners had perceived imitation (i.e., the intercept was positive and significant in both samples). In the remaining 22 runs, either the first or second sample had an insignificant intercept. Since 97.8% of the runs resulted in samples that reached the same conclusion, this provides further evidence of the reliability of the AXB task.

4.1.2. Reliability across different numbers of listeners

Across studies using the AXB task to assess imitation, the number of listeners per shadower varies widely from four or five per shadower (Pardo et al., 2013; Pardo et al. 2017; Pardo et al., 2018) to 25 per shadower (Clopper & Dossey, 2020). How might the number of listeners who respond to each shadower influence the reliability of the findings? Pardo et al. (2017) assessed reliability of AXB responses by comparing split-halves of responses from a first and second block of the AXB test and comparing the mean and standard deviation of %-correct of 10 listeners, compared to five, four, three, two, and one listeners. They found that the mean and standard deviation of %-correct were similar for 10 listeners as compared to five, four, and three listeners, but reliability dropped when there were fewer than three listeners. This method focuses on %-correct and does not tell us whether the outcome of a statistical model would be the same as the number of listeners drops. As such, another way to test reliability of the findings from an AXB task is to compare the overall finding of "significant imitation" at the group level, comparing groups of listeners of different sizes. In the current study, with a total of 40 listeners evaluating all the shadowers, the overall finding was that the shadowers imitated (i.e., the intercept was significant). Would we come to the same conclusion using fewer than 40 listeners? What is the minimum number of listeners needed to reach the same conclusion as we do with 40 listeners? We explored this by making random groups of 2 to 39 listeners and using the subset of AXB responses within each group as the dataset for the model in (1), which tests whether the overall finding is imitation. We took 40 unique random samples for each group size.[9] **Figure 1** below shows the relationship between the group size (on the x axis) and the proportion of samples within each group size for which the intercept in the model was significant (on the y axis), mirroring the finding using all listeners.

The plot shows that 100% of the samples found overall imitation when group size was between 39 and 20. Once group size drops below 20, we start to see a reduction in consistency of the findings of the samples. From around 20 to 6 listeners per group, there is a more or less linear decrease in the proportion of samples for which the intercept was insignificant (i.e., a finding of no imitation), down to around 80%. Below 6 listeners, the consistency falls sharply, ending up at roughly 50% of samples finding imitation for a group size of 2. This suggests that, at least for the current dataset, reliability of the finding does not increase beyond 20 listeners, but that below 20 listeners, reliability of the findings drops, with a substantial drop below 6 listeners. This pattern is somewhat different from that reported by Pardo et al. (2017), which found that consistency of the finding was strong down to three listeners. However, the approach used in that study compared the mean and standard deviation of %-correct and did not consider whether

[9] The choice to generate 40 samples per group size is based on the fact that the maximum number of possible samples of size 39 that can be generated from a total of 40 possibilities is 40.

the samples generated would result in a statistically significant finding of imitation. Arguably, doing so is more important than only examining %-correct since it is through the model that we evaluate whether or not a group of participants have imitated (at least when using a perceptual assessment method).

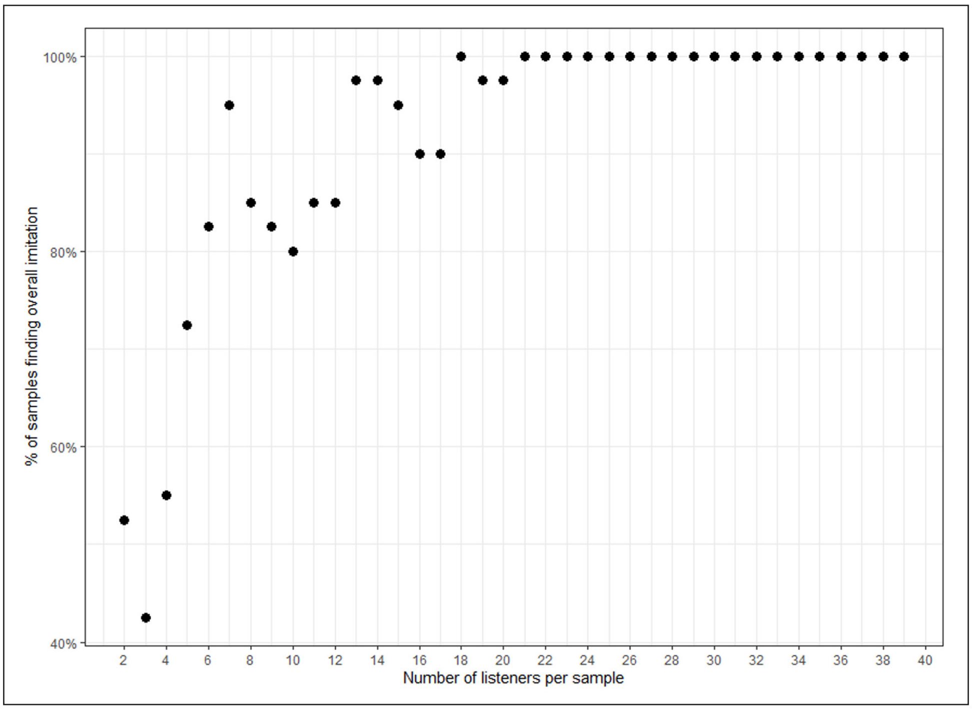

Figure 1: Percentage of samples in which intercept is significant (i.e., imitation found) by number of listeners per sample.

Taken together, the results of these analyses indicate that the AXB task is a reliable tool to assess imitation as long as enough listeners are included. The reliability of the findings stabilizes at around 20 listeners with reliability remaining high (above 80%) down to six listeners. Note, however, that this finding is likely highly dependent on the number of tokens, how much the shadowers imitated the model talkers, the strength of the effect needed, and various other factors.

4.2. Extent and consistency of individual variation in perception

Next, I consider variation in the perception of imitation at the individual listener level. **Figure 2** plots the distribution of the individual %-corrects. The red dashed line shows the mean %-correct across listeners (58.8%). At the lowest end, there are listeners whose %-correct is close to 50% (lowest is 51.4%), meaning that they are choosing the shadowed and baseline tokens roughly equally. This could suggest that they are guessing. Guessing could be the result of not paying attention or hitting responses randomly, or it could mean that those listeners are genuinely not able to perceive imitation, either because they lack the necessary perceptual sensitivity or because there was no actual imitation in the recordings to perceive. In this study, it is unlikely that there was genuinely no imitation to perceive since other listeners' %-correct was much higher (highest is 69.9%) and all participants completed the same experiment. A %-correct of 65% or higher is well above those typically reported in AXB studies of imitation, where the group-level mean generally falls between 52% and 58% (Kim, 2012; MacLeod & Di Lonardo Burr, 2022; Nielsen & Scarborough, 2019; Pardo et al., 2017; Shockley et al., 2004; Wagner et al., 2021; Walker & Campbell-Kibler, 2015). This suggests that there is substantial variation between listeners in how likely they are to choose the shadowed token as being more similar to the model talker. In other words, it seems that listeners vary in their ability to perceive imitation.

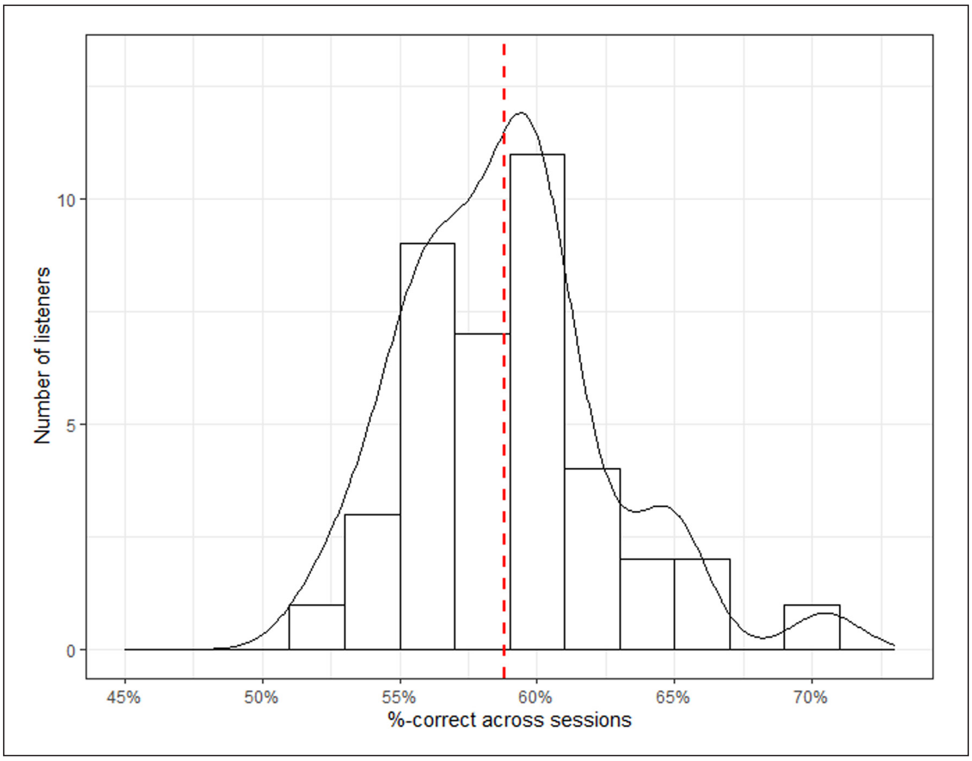

Figure 2: Histogram and density plots of percent correct by individual listener for both sessions together. The red dashed line shows the mean percent correct.

To determine if this variation is statistically significant, we can compare the model given above in (1) to one without the random intercept and slope for listener using likelihood ratio testing. This new model is given in (2). The random intercept for listener captures individual variation in the likelihood of choosing the shadowed token in each session. If including that term improves model fit, this suggests that there is significant variation in %-correct among the listeners.

(2) glmer (CorrectYN ~ 1 + SESSION + AXB ORDER + EXPOSURE + (1|shadower) + (1|word)

The model that includes the random intercept and slope for listener does improve model fit over the one without it ($\chi^2(3) = 899.23$, $p < 0.001$). This provides statistical support for the impression given in **Figure 2** that the listeners vary in how often they choose the shadowed token.

Next, we want to know how consistent that variation in listeners is across sessions. While a range of %-corrects was found, it could be that the variation is random and would not be repeated in the second session. If the %-corrects of the individuals correlate across sessions, this suggests that the differences in the ability to perceive imitation among the listeners reflect a stable ability of the listeners, rather than random variation. **Figure 3** plots the %-correct for each listener in Session 1 on the x-axis and Session 2 on the y-axis. If the proportion of trials in which the listeners choose the shadowed token is stable across sessions, then we should expect a positive correlation between Session 1 and Session 2.

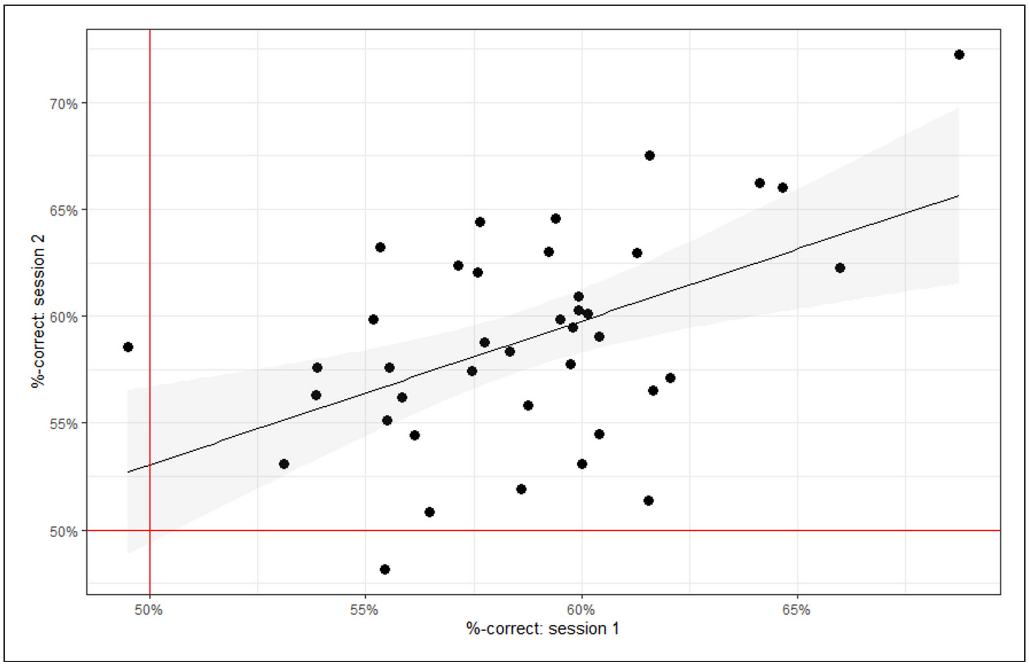

Figure 3: Percentages of shadowed tokens chosen in Session 1 by Session 2 for each individual listener.

Figure 3 shows that there is a general positive relationship between the individual %-corrects across sessions. As discussed earlier, a standard method of assessing test-retest reliability is with Pearson correlations (Bannigan & Watson, 2009; Carmines & Zeller, 1979, p. 38; DeVon et al., 2007). Applying that approach here reveals that the %-corrects across sessions for the individual listeners are significantly correlated ($r(38) = 0.50, p < 0.01$), indicating that there is some degree of reliability in the performance of the listeners across sessions. This suggests that the variation in %-correct observed in **Figure 2** is not simply random variation but might reflect differences in the ability of the listeners to perceive imitation – differences which persist across sessions.

Next, we can consider how this consistency of the %-correct scores might differ when we look at the scores for each of the four shadowers separately. **Figure 4** shows the relationship between the %-correct scores in Session 1 and Session 2, with each shadower in a separate panel. We can see that for each shadower there is a generally positive relationship between the sessions; however, the slope of the regression line in each differs. For shadower A, the slope appears to be steepest, while for shadower D, it is flattest. Indeed, if we explore the correlations for each shadower, we find that shadower A has the highest correlation in scores between the two sessions ($r(38) = 0.62, p < 0.001$), followed by shadower B ($r(38) = 0.55, p < 0.001$), with the scores for the other two shadowers not being significantly correlated (shadower C: $r(38) = 0.27$, $p = 0.09$; shadower D: $r(38) = 0.19, p = 0.22$).

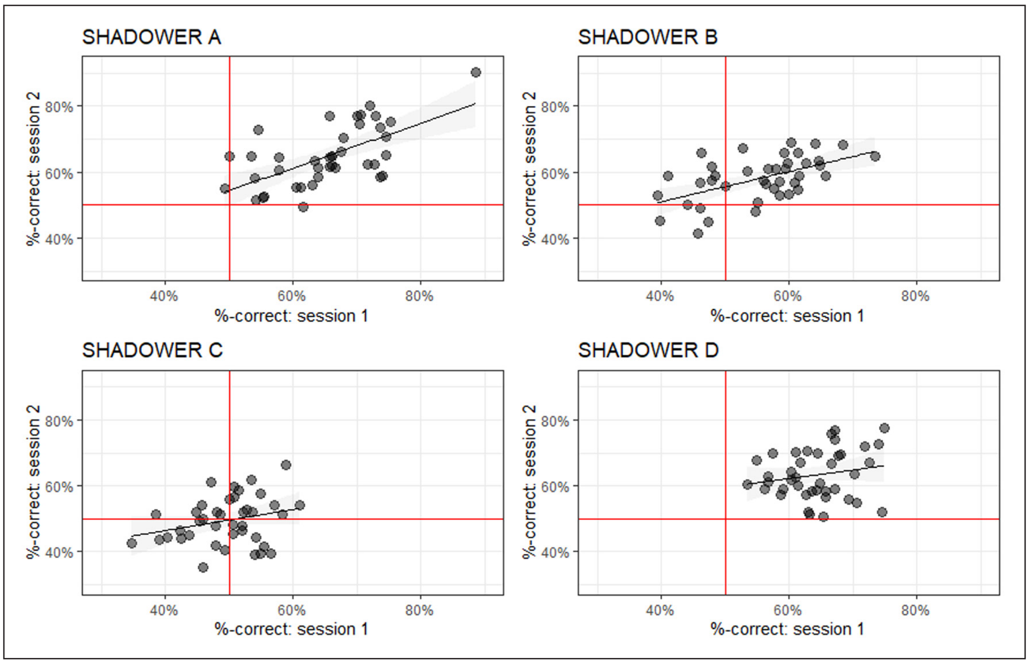

Figure 4: Percentages of shadowed tokens chosen in Session 1 by Session 2 for each individual listener split by shadower.

This indicates that listeners differ in how consistent their %-correct scores were depending on the shadower they listened to. Note that the %-corrects for shadower C are centred around 50% in both sessions. Applying how the AXB task is used to measure imitation suggests that shadower C did not imitate the model talker much, creating little imitation for the listeners to respond to. If shadower C did not imitate or diverge much, then listeners would have little reason to choose either of the shadowed or baseline tokens as being more similar to the model talker and so we might expect a %-correct of around 50%. In that case, there is no reason to expect listener responses to be consistent across the sessions, and that is what we observe here. This does not apply to shadower D, however, whose %-corrects are generally above 50% in both sessions. In comparison to shadowers A and B, however, shadower D's %-corrects in Session 1 are more tightly clustered around the means. As such, shadower D has a narrower range of %-corrects in Session 1 than the other two. Perhaps this narrower range prevents the correlation between %-correct in Session 1 and Session 2 from being significant, while the broader ranges of %-correct for shadowers A and B allows the correlation to be significant.

4.2.1. Consistency of trial-by-trial choices

To this point I have considered the percentage of trials in which the listeners select the shadowed token as being more similar to the model talker production. While this approach reveals whether the %-correct is related across sessions, it does not determine how consistent listeners are on a trial-by-trial basis. While participants might have a similar %-correct across sessions, it could be that the specific trials they choose the shadowed token on are quite different. In this section, I explore to what extent the Session 1 choices predict the Session 2 choices. Put another way, how likely are the listeners to make the same choice for the same trial in both sessions? As noted in §3.2.2, in each session the listeners responded to 320 trials. To test the trial-by-trial consistency, I consider how often listeners made the same choice on those 320 trials in Session 1 and in Session 2.

This behaviour can be explored by considering the proportion of trials in which they chose the shadowed token in both sessions or the baseline token in both sessions (which will be referred to as %-same). As a group, the 40 listeners made the same choice on 61.4% of trials. The distribution is shown in **Figure 5** below. The red dashed line shows the mean %-same. Listeners ranged in their %-same from 50% to 74%. Those with %-same close to 50% only made the same choice about half the time and could be considered very inconsistent. Note that there is only one participant at the 50% mark, but several more only a bit higher than 50%. On the other hand, there are participants making the same choice in 65% or more of trials. These participants could be considered much more consistent in their trial-by-trial choices. These findings suggest that listeners vary in how consistent they are, but note that no participant has a %-same below 50%.

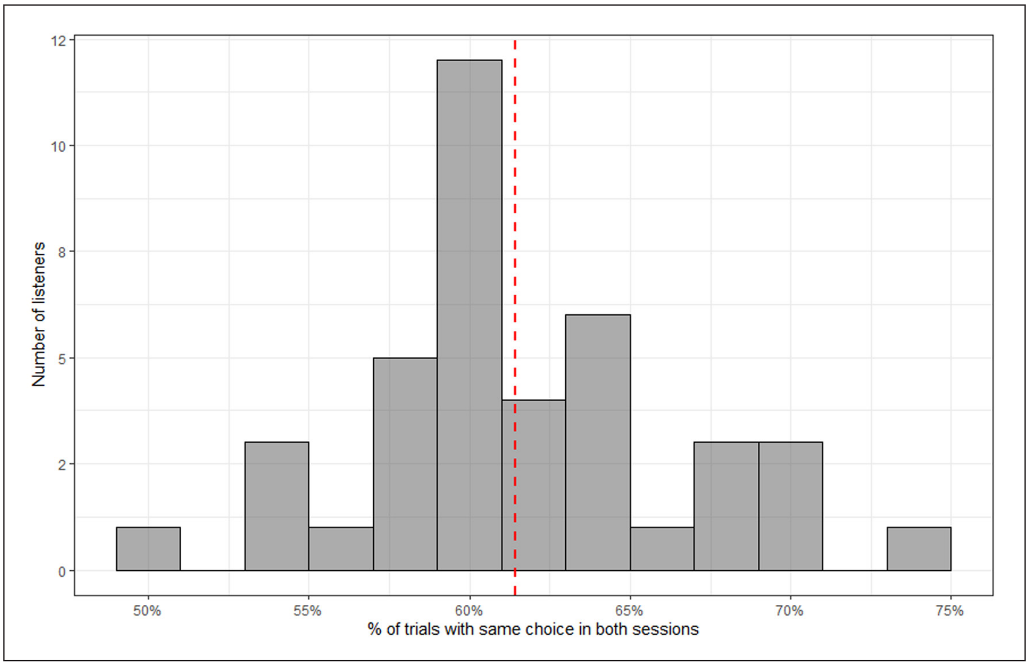

Figure 5: Histogram of proportion of trials on which listeners made the same choice in both sessions (%-same).

To confirm that the mean %-same of 61.4% is higher than 50%, a model was built with a dependent variable of SAMEYN with two levels (yes = same choice in both sessions; no = different choice between sessions), fixed effects of EXPOSURE and AXB ORDER, and random intercepts for listener, shadower, and word.[10] The model also includes a by-listener slope for AXB ORDER and is given in (3) below.

(3) $$\text{SAMEYN} \sim 1 + \text{EXPOSURE} + \text{AXB ORDER} + (1 + \text{AXB ORDER}|\text{listener}) + (1|\text{shadower}) + (1|\text{word})$$

The intercept was positive and significant ($\beta = 0.478, p < 0.001$), as shown in **Table 3**, indicating that 61.4% is statistically significantly higher than 50%. AXB order was significant, showing that when the shadowed token is in the A position, the listeners were more consistent. However, the estimate is relatively small (0.08) and the difference in %-same is quite small: 60% when the shadowed token is in the B position vs. 62% when in the A position.

[10] A reviewer asks how consistency of the responses might differ among the 20 stimuli. The %-same for the stimuli ranges from 52% to 82%, but 15 of the 20 words have a %-same between 55% and 65%. Comparing the model in (3) to one without the random intercept for word finds that having the intercept provides a better model fit (χ2(1) = 130.49, p < 0.001). This suggests that there is significant variation among the words in terms of how consistently the listeners responded to them across the sessions.

	β	**SE**	**p**
Intercept	0.48	0.09	< 0.001
EXPOSURE	0.07	0.04	0.0929
AXB ORDER	0.09	0.05	0.0750

Table 3: Statistical results from a logistic mixed effects regression to determine if listeners' proportion of consistent responses is statistically significant.

It is worth considering how the nature of the choice might influence consistency. It is a central assumption of the perceptual method of assessing imitation that if there is no imitation to be perceived, then the listeners will choose the shadowed and baseline tokens as being more similar to the model talker roughly equally. In that situation, we would expect little consistency across two sessions: given a particular choice in Session 1, listeners would be equally likely to make the same or opposite choice in Session 2. On the other hand, if there is imitation to be perceived, listeners should choose the shadowed token more often than the baseline overall. Furthermore, on trials that contain imitation, listeners should be more likely to choose the shadowed token again in the second session. The above results suggest that there is, in fact, imitation to be perceived, at least on some of the trials. As such, we might expect to find an asymmetry between choices in the second session that depends on the choices in the first. When the listeners choose the shadowed token in the first session (i.e., they perceive imitation), they should be more likely to do so again in the second session than if they choose the baseline token in the first session.

In the current data, when listeners chose the shadowed token in the first session, they made the same choice on 67% of trials in the second session; but, when they chose the baseline token in the first session, they only made the same choice on 53% of trials. **Figure 6** plots the distribution of these %-same proportions for the 40 individual listeners, split by whether they chose the baseline or the shadowed token in the first session. The 50% mark is illustrated using a red dashed line. The closer the %-same distribution is to this line, the less of a relationship there is between the Session 1 and Session 2 choices. The higher the %-same, the more often the listeners make the same choice in both sessions.[11] We can see that the distribution of %-same when they chose the shadowed token in the first session is higher than that of when they chose the baseline token, with a small amount of overlap between the distributions.

To determine to what extent the Session 1 choice influences consistency on a trial-by-trial basis, we can add in the Session 1 choice as a predictor to the model given in (3), generating the model in (4), which includes by-listener slopes for AXB ORDER and Session 1 choice.

[11] If the distribution were below 50%, this would indicate that the listeners were consistently changing their choice from Session 1 to Session 2.

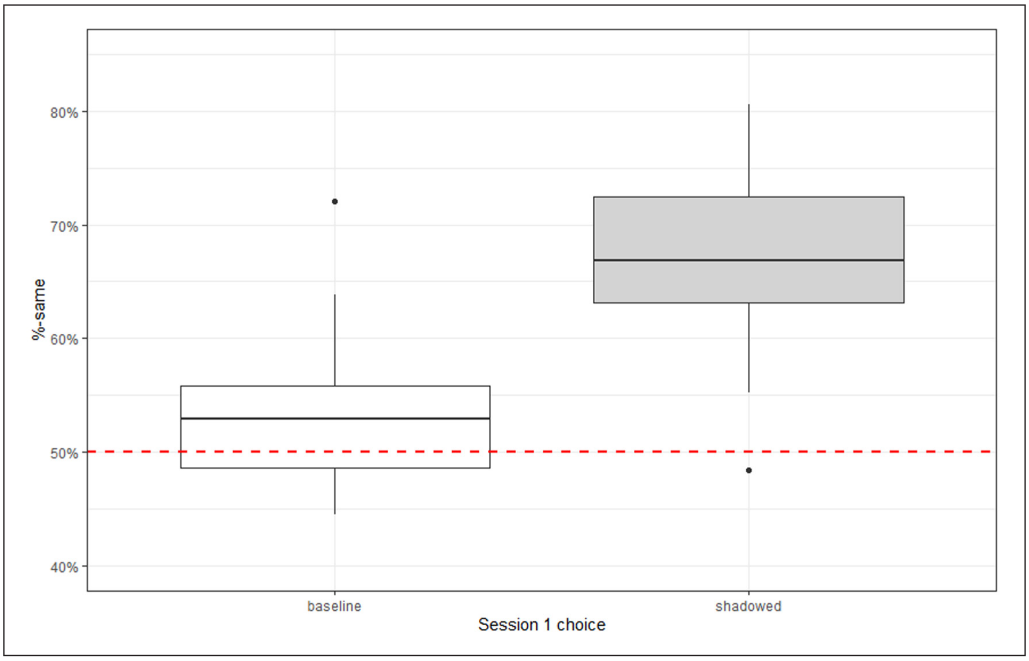

Figure 6: Boxplot of proportion of trials on which listeners made the same choice in both sessions, split by whether they chose the baseline or shadowed token in Session 1. Dashed line shows the 50% mark.

(4) $$\begin{aligned}\text{SameYN} \sim 1 + &\ \textsc{Session1\_choice} + \textsc{exposure} + \textsc{axb order} \\ &+ (1 + \textsc{axb order} + \textsc{Session1\_choice}|\text{listener}) + (1|\text{shadower}) \\ &+ (1|\text{word})\end{aligned}$$

The output is provided in **Table 4**. The results show that the choice of shadowed or baseline in Session 1 is highly related to the consistency of the choices at the group level. The likelihood of making the same choice in both sessions increases significantly when the Session 1 choice is the shadowed token as compared to when it is the baseline token[12].

This finding that listeners are more consistent when they choose the shadowed token in Session 1 generally holds up when we look at the four shadowers separately as well. **Figure 7** shows that for three of the four shadowers, the %-same is higher when the listeners chose the shadowed token in the first session than when they chose the baseline token. The exception is

[12] A reviewer suggests that if the Session 1 choice influences the consistency of responses across sessions, it should also be true that, within each session, the Exposure 1 choice should influence the consistency of responses across exposures. To test this, I built two models (one per session), similar to the one in (4), exploring the extent to which the Exposure 1 choice (shadowed vs. baseline) influenced the percentage of trials on which listeners made the same choice on the first and second exposure. For both sessions, the Exposure 1 choice is significantly related to consistency across exposures; when listeners chose the shadowed token in Exposure 1, this significantly increased the consistency across exposures (Session 1: $\beta = 0.92$, $p < 0.001$; Session 2: $\beta = 1.42$, $p < 0.001$).

	β	SE	p
Intercept	0.44	0.08	< 0.001
Session 1 choice: baseline vs. shadowed	0.55	0.06	< 0.001
exposure	0.07	0.04	0.1125
AXB order	0.07	0.04	0.0592

Table 4: Statistical results from a logistic mixed effects regression predicting consistency by Session 1 choice.

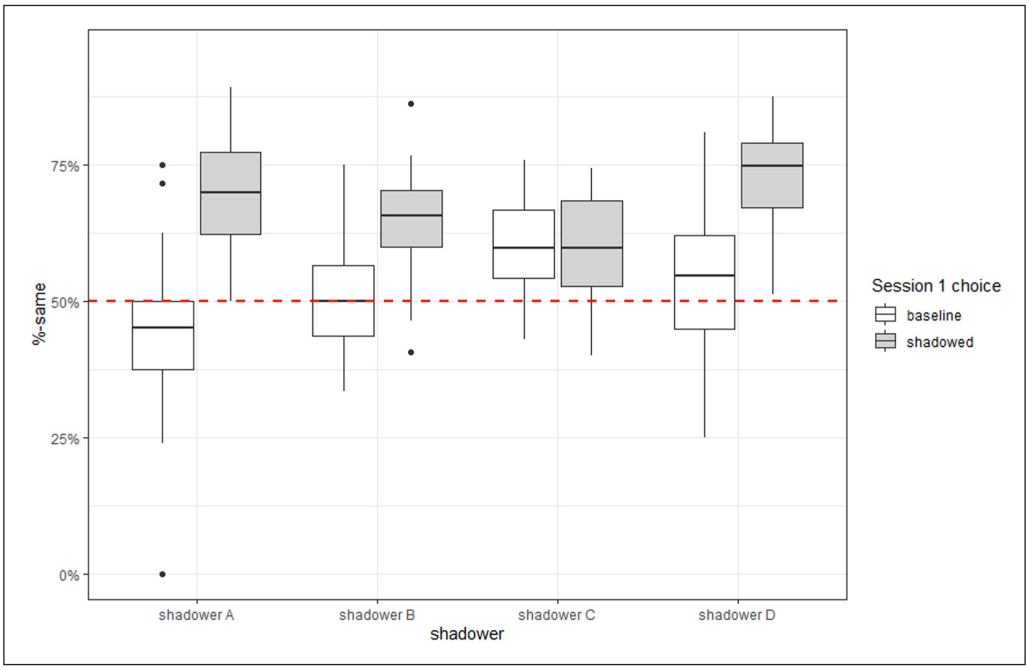

Figure 7: Boxplot of proportion of trials on which listeners made the same choice in both sessions, split by whether they chose the baseline or shadowed token in Session 1 and by shadower. Red dashed line shows the 50% mark. Shadower C, for whom listeners are almost exactly as consistent regardless of their Session 1 choice.

This pattern holds at the individual level as well. **Figure 8** shows the proportion of trials on which each listener made the same choice, split by whether they chose the baseline or shadowed token in Session 1. The grey dots show the proportion of trials on which each listener chose the shadowed token in the first session and then made the same choice in the second session. The white dots show the proportion of trials on which each listener chose the baseline token in the first session and then made the same choice in the second session. The participants are ordered by the magnitude of the difference in this %-same between trials in which they chose the shadowed and trials in which they chose the baseline in the first session. That is, participants at the left of the plot make a large difference in consistency depending on whether they chose the shadowed token or the baseline, while those on the right of the plot show about the same consistency regardless of their Session 1 choice. Notice that for almost all participants (36/40), the proportion of trials in which they make the same choice in both sessions is higher when they choose the shadowed token in Session 1 than when they choose the baseline token. In many cases, this difference is quite large. This shows that if the participants choose the shadowed token in the first session, they are much more likely to choose it in the second session than if they choose the baseline token in the first session. In fact, when they are incorrect in the first session, 29 of the 40 participants are correct no more than 55% of the time in the second session. In contrast, only one participant is correct less than 55% of the time in the second session after choosing the shadowed token in the first.

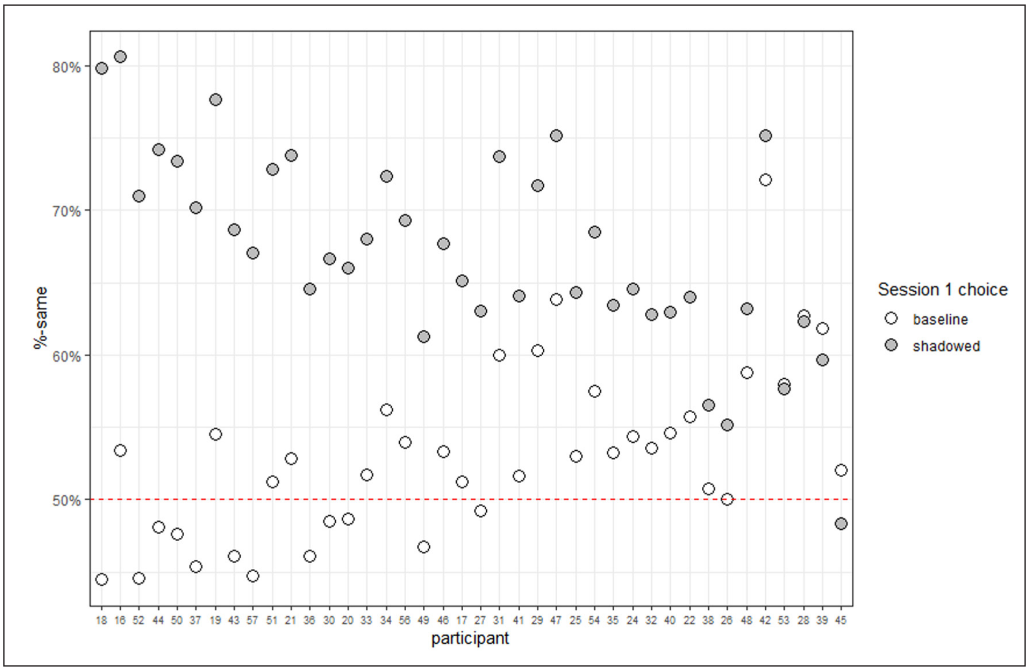

Figure 8: Scatterplot showing the proportion of trials in which listeners made the same choicein both sessions split by whether they chose the shadowed token (grey) or baseline token (white) in Session 1.

This section showed that the listeners' choices were related across sessions, with the choice they made in Session 1 being highly predictive of the choice they made in Session 2. However, the extent to which the Session 1 choice was predictive depending heavily on what that choice was. When they chose the shadowed token in Session 1, they were very likely to choose it again in Session 2, but when they chose the baseline token in Session 1, their Session 2 choices were much less strongly related to the Session 1 choices. This finding is discussed further in section 5.2.

5. Discussion

The goals of this study were to examine the reliability of the AXB task as an assessment of phonetic imitation and to explore the extent and consistency of individual variation in the perception of imitation. Relating to the first goal, the overall proportion of trials on which the listeners chose the shadowed token was almost identical in both sessions, and both proportions were significantly higher than 50%. Following how the AXB task is typically used in studies of phonetic imitation, these results allow us to conclude that the shadowers imitated, regardless of whether Session 1 or Session 2 responses are used. There were also differences in how much imitation was perceived among the four shadowers and those differences were also consistent across sessions. Taken together, these results indicate that the AXB task is a reliable tool for measuring phonetic imitation when considering all of the listeners as a group.

Relating to the second goal, the results showed that listeners ranged in their overall %-correct from quite close to 50% to proportions much higher than those typically reported at the group level (around 70%). Furthermore, the listeners were fairly consistent in their individual %-corrects across sessions, suggesting that the variation in ability is not simply random, but rather a stable property of the individual participants' ability to perceive imitation. However, while there was a significant positive correlation in individual %-corrects across sessions, there was also a substantial amount of variability in how consistent the scores were across sessions. Part of this variability could be due to differences in individual consistency in %-correct depending on which shadower the listeners were responding; the listeners' %-corrects were found to be significantly correlated for two of the shadowers (Shadowers A and B), but not the other two (Shadowers C and D). Focusing on consistency of the trial-by-trial choices revealed that the Session 1 choices were strongly related to the Session 2 choices, but that how strongly they were related depended on what the Session 1 choice was. When the listeners chose the shadowed token in the first session, they were much more consistent in making that same choice in the second session than when they chose the baseline token in the first session. This pattern persisted at the individual shadower[13] and individual listener level.[14]

Taking the results together, one of the main findings is that listeners differ in their ability to perceive phonetic imitation in the AXB task. Individual variation in perception has been observed in a variety of other lines of research including use of cue weights in perception of contrasts (e.g., Kong & Edwards, 2016; Schertz et al., 2015), cross-language perception (e.g.,

[13] Except within Shadower C, for whom the listeners' consistency was almost exactly the same for each Session 1 choice.

[14] 36/40 listeners showed this pattern.

Harnsberger, 2000), perception of second language contrasts (e.g., Kim et al., 2017; Mayr & Escudero, 2010), and perception of speaker-specific phonetic detail (see Smith, 2015 for a review). What might cause this variation to exist? According to Yu & Zellou (2019), figuring out why this variation exists is "a largely unsolved puzzle" (Yu & Zellou, 2019, p. 133), but several possible contributors have been advanced. One is that individuals might vary in their perceptual sensitivity to different acoustic-phonetic dimensions (Wade et al., 2020), making some "better phonetic perceivers" (Hall-Lew et al., 2021, p. 9) than others. This idea is similar to the notion of phonetic talent explored by Lewandowski and Jilka (2019), where talent is defined as "a stable, innate characteristic that is separate from external circumstances of learning" (p. 5). Social and cognitive characteristics of the individual and details of the individual's experience could contribute to variation in speech perception as well, including amount of music experience (e.g., Chandrasekaran et al., 2009), individual patterns in attention and memory (Cowan et al., 2005; Yu & Zellou, 2019: 140), inhibitory skills (Darcy et al., 2016; Lev-Ari & Peperkamp, 2014), executive function (e.g., Kong & Edwards, 2016), social network size (Lev-Ari, 2018) and degree of autistic-like traits (Yu & Zellou, 2019: 140, Stewart & Ota, 2008). A reviewer points out that in addition to factors such as those discussed above, which could stem from differences in perceptual acuity generally, some of the variation in the perception of imitation could also result from differences in the particular acoustic-phonetic characteristics that listeners attend to when perceiving imitation. It is likely the case that variation in the perception of imitation is caused by both types of factors.

The above studies show that individual variation in perception is common and attributable to a wide range of factors. The results of the current study indicate that listeners also differ in their ability to perceive phonetic imitation. Future work will need to be carefully designed to determine which factors explain that variation.

5.1. Implications for our use of the AXB task

What do the current findings mean for our use of the AXB task to assess phonetic imitation? Since listeners do vary in their ability to perceive imitation, this means that we cannot assume that a group-level mean %-correct from an AXB task will accurately reflect the abilities of the individuals who completed the task. In other words, we cannot assume that the individuals will perform similarly to each other. Indeed, as observed in this study, listeners ranged in their %-correct from 51.4% to 69.9%, despite all listeners evaluating the same trials. If a group-level %-correct were close to 50%, we would take this as evidence that the shadowers did not imitate. In contrast, if the group-level %-correct were close to 70%, that would be very strong evidence of there being imitation in the signal. Since many of the listeners have %-corrects well above 50%, this suggests that there is, in fact, imitation available to be perceived. What can be different among the listeners, however, is their ability to perceive that imitation. The current study suggests that this variation does exist and that it is a relatively stable property of individuals, since the listeners' %-corrects were correlated across sessions.

This variation has some practical implications for our use of the AXB task to assess imitation. In many studies, there are many shadowers and/or tokens that need to be assessed by listeners. Because of this, listeners often only assess a subset of the total set of trials (e.g., Pardo et al., 2017). As a result, different listeners assess different shadowers or words. Given that listeners vary in their performance in the AXB task, having listeners evaluate different subsets of shadowers or words potentially leads to a confound: If we find a difference between shadowers, we cannot be sure whether the difference is the result of the shadowers having imitated to different degrees or if it is due to differences among the listeners' perceptual abilities in detecting imitation, or both. However, the extent to which this confound is an issue could depend on the number of listeners assigned to each shadower. The current study (in section 4.1.1) tested whether splitting up the listeners among the four shadowers would change the overall conclusion that the shadowers imitated. The findings indicated that of the 1000 random groupings of 10 listeners assigned to one shadower, 88.5% reached the same conclusion as the full dataset, where imitation was detected with no difference between the sessions. It seems then that 10 listeners per shadower are likely sufficient. Even so, in studies that have a large number of tokens, asking listeners to respond to vast numbers of trials is not feasible. One way to avoid this problem is to include a smaller subset of trials that are common to each listener and then to compare listener performance on that subset to determine the degree of inter-listener variation before comparing trials that differ between listeners. At the very least, this approach would allow us to find out the extent of the individual variation in a particular study.

The fact that listeners vary in their ability to perceive imitation also suggests that we should ensure that we include enough listeners in our AXB studies (as previously suggested by Pardo, 2013) to avoid inadvertently skewing our findings by disproportionately including listeners who tend towards one end of the range of abilities. The more listeners we have, the less likely this skewing is to happen. However, the results of the analysis in section 4.1.2 suggested that there might be diminishing returns when including more than 20 listeners, at least for a study with a similar number of shadowers and tokens as the current study. On the other hand, when the number of listeners dropped below six, the overall finding (of imitation or not) became much less reliable. Taken together, our findings suggest that including 10 listeners per shadower is likely sufficient.

That listeners vary in their perception of imitation also suggests that we should be reporting more about the listeners in our studies in two main ways. For one, we should report variation in %-correct so that we can continue to amass evidence regarding the extent to which listeners vary in their ability to perceive imitation. Also, we should include more information about the listeners' demographic characteristics (such as age, gender, region of origin, race, language experience, etc.). Doing so allows us to gain a better idea of how such characteristics might be related to variation in the perception of imitation. Even if the goal of the study is not to explore these things, including this information also normalizes doing so and helps us ensure that we are collectively testing as broad a set of participants as possible.

As a whole, these results suggest that the AXB task is reliable as a tool for assessing imitation as long as we include enough listeners per shadower. Furthermore, if listeners only assess a subset of trials, we need to be cautious about interpreting the influence of factors (such as dialect or gender of shadowers) that might differ between those subsets.

5.2. Interpretation of choosing a baseline versus shadowed token in the AXB task

Section 4.2.1 considered how the consistency of choices from Session 1 to Session 2 might be influenced by which choice the listeners made in Session 1. The findings provide an opportunity to focus on the expectations of how the AXB task will work and how we can interpret results. First, as noted earlier, we assume that if the shadowers did not imitate the model talker, then there will be no imitation to be perceived, and in that case, we would expect the listeners, who are required to make a choice, to choose the baseline and shadowed tokens roughly equally across trials.[15] We take this as a kind of null hypothesis and compare our AXB findings against this backdrop. If listeners choose the shadowed token more than 50% of the time, we take this as evidence that they are perceiving imitation.

From here, we can extend these predictions to include consistency of listener responses across sessions (or across exposures to the same trial within a session). On trials that contain imitation, we might expect listeners to perceive that imitation in both sessions and therefore choose the shadowed token in both sessions, leading to high consistency. On the other hand, for trials that do not contain imitation, we fall back to the basic prediction of the AXB task and expect that in Session 1, there should be a roughly 50/50 split of baseline versus shadowed responses. Although this is not often specifically stated, the implication of this expectation is that, in the absence of imitation, listeners are guessing. Similarly, in Session 2, there should also be a 50/50 split; but, given that listeners are guessing again, we should not expect a relationship between the Session 1 and Session 2 choices; in that case, consistency should be low.

The analysis of consistency in section 4.2.1 showed that when the listeners chose the shadowed token in Session 1, they made the same choice on 67% of trials in Session 2. In contrast, when they chose the baseline token in Session 1, they made the same choice on only 53% of trials. A statistical model confirmed that having chosen the shadowed token in the first

[15] Note that such a finding could be the result of other factors such as a listener's lack of perceptual acuity or attention or a genuine combination of convergence and divergence among the shadowed tokens.

session significantly increased consistency across the sessions. **Figures 7** and **8** illustrated that this pattern held up across shadowers and across listeners. These results align with the predictions made above and suggest that when there is some imitation to perceive, listeners are likely to perceive it both times. When there isn't, listeners might guess both times, leading to a roughly 50/50 split of making the same choice versus different choice across sessions.[16] A third possibility is that a shadower diverges from the model talker. In that case, the listeners will be more likely to perceive the baseline token as being more similar to the model talker. In such a case, we would expect listeners to be more likely to make the same choice in the second session, again leading to (relatively) high consistency. It is possible that this occurred for some of the tokens for Shadower C. As shown in **Figure 7**, consistency across sessions was above 50% and did not seem to depend on the listeners' Session 1 choice when responding to Shadower C.

It is important to note that the interpretation of AXB responses requires being applied across many trials; it cannot be applied on individual instances of individual tokens. For example, we cannot look at a specific instance of the word 'cyclist' and determine that imitation occurred simply because a listener selected the shadowed token as being more similar to the baseline on one trial. Similarly, the choice of 'baseline' on an individual token does not necessarily mean that the listener was guessing or that there was no imitation in that token. Instead, we look across distributions of tokens to see how the patterning compares to our null hypothesis and expectations about consistency.

5.3. Communication Accommodation Theory

One of the main theoretical positions typically taken on the mechanism behind or motivation for phonetic imitation holds that the reason talkers imitate is social. As discussed in section 1, Communication Accommodation Theory (CAT: Giles, 1973) posits that talkers in an interaction converge towards an interlocutor to minimize social distance between themselves and their interlocutors. Research has shown that there are social consequences to imitating (or not). For example, people who imitate have been shown to be perceived as more likeable and attractive (e.g., Chartrand & Bargh, 1999) and conversations in which imitation takes place have been found to be rated more positively than those in which imitation does not occur (Giles & Smith, 1979; Street, 1982). Crucially, CAT predicts that talkers imitate to influence their interlocutor's perception of them and their interaction. However, this influence can only take place if the interlocutor can perceive that imitation has been produced. As Dias et al. (2021) note, "if phonetic

[16] Note that it is difficult to say for sure which trials contain imitation and which do not. It is the purpose of the AXB task to determine that, but more at a global level, rather that for individual trials. Acoustic analysis can also be used to measure imitation too, but again, is typically calculated across many trials. Stimuli that were modified along some acoustic dimension (such as VOT) provide an opportunity to test accuracy and consistency of the perception of imitation on trials that we know a priori contain (or do not contain) imitation.

convergence does have some social relevance, as some have argued, then its existence should be *perceptible*" (Dias et al. 2021, p. 1, emphasis in original). The findings of the current study suggest that listeners vary in the ability to perceive imitation. This could mean that individuals will vary in the extent to which they can capitalize on the social information encoded in phonetic imitation. If so, this might reflect variation in listeners' ability to access social cues, both in their own conversations and in conversations they observe between others. Listeners who are better able to detect imitation might be better able to evaluate rapport between interlocutors (e.g., Pardo et al., 2012) and determine social relationships and identify social leaders (e.g., Dias et al., 2021; Giles et al., 1991; Pardo et al., 2012; Shepard et al., 2001). The current study finds that listeners do vary in detecting imitation; future work could explore how that variation might relate to differences in accessing social information about other speakers and conversations. However, to do so, we would need to explore the perception of imitation in conversation as well. Just as previous work has shown that imitation in a shadowing task only weakly correlates with imitation in conversation (Pardo et al., 2018), it could be that the ability to perceive imitation in an AXB task does not directly reflect the ability to perceive it in conversation.

5.4. Possible episodic effects

To this point, I have considered the fact that the listener's performance in the AXB task is related across sessions to be evidence of stable individual-level differences in ability to perceive imitation. However, this relationship could also be caused by episodic effects[17], where the listeners retained some memory of the individual tokens and their realization from Session 1 and then accessed them in Session 2. The possibility of episodic or memory effects represents one of the limitations of the test-retest paradigm. According to Carmines & Zeller (1979, p. 38–39), participants retaining some memory of the test from the first session can cause us to overestimate the reliability of our empirical measurements. How much time would need to elapse to preclude episodic effects is an open question, but if we assume that episodic effects are more likely when the two sessions are close together, then we might expect that listeners who completed Session 2 after the least number of days might be more consistent in their responses than participants who waited longer between sessions. As mentioned in section 3.2.2, this amount of time ranged from 12 to 27 days, with a median of 16 days. **Figure 9** explores the relationship between the individuals' %-same and the number of days between their sessions. The left-hand panel shows a weakly negative relationship, but the two measures are not significantly correlated (r(38) = –0.23, p = 0.1634). However, recall that the analysis in section 4.2.1 found that the listeners' consistency (i.e., %-same) across the sessions was strongly influenced by the choice they made in the first session, with consistency being much higher when they chose the shadowed token as

[17] Thank you to an anonymous LabPhon conference abstract reviewer for this suggestion.

being more similar to the model talker in Session 1. If the plot in **Figure 9** is split by the Session 1 choice (given in the right-hand panel), we find that when the listeners choose the shadowed token in Session 1 (grey dots), their consistency is more strongly related to the number of days between sessions than when they choose the baseline token (white dots). In fact, %-same and number of days between sessions is significantly negatively correlated when they choose the shadowed token (r(38) = –0.37, p < 0.05), but the relationship is not significant when they choose the baseline token (r(38) = 0.17, p = 0.3052). This might mean that when listeners perceive imitation in Session 1, they are better able to retain some memory of that experience in Session 2 the closer the two sessions are to each other. In contrast, when they do not perceive imitation (i.e., they choose the baseline token), the number of days between sessions has no influence on the likelihood of making the same choice, suggesting that they retain no memory of that experience.

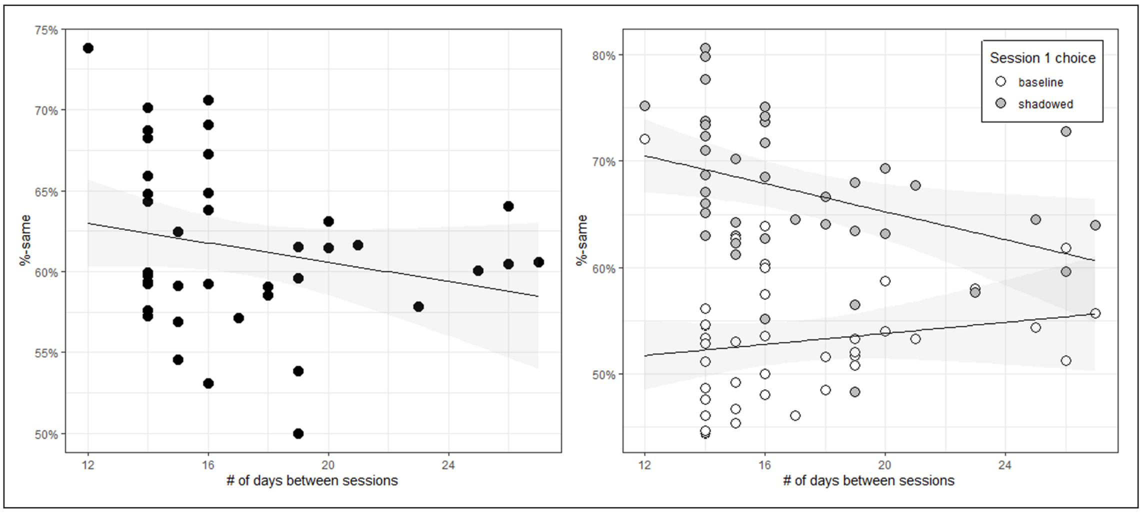

Figure 9: Relationship between %-same and days between sessions (left) and split by Session 1 response (right).

On the other hand, it could be that participants who sign up for the earliest second session (closer to 12 days) are more focused or interested in the experiment and, as a result, are more consistent. To know for sure, we would have to ensure that the interval between sessions was random among the listeners and not decided by them, as it was in this study. In addition, there are a lot more data points towards the shorter end of the possible intervals between sessions than there are at the longer end, with quite a bit of variation in %-same even among those who completed the second session after only 14 days. Taken together, these results suggest that episodic effects are possible, but cannot be teased apart from other influences. To the extent that episodic effects are at play in the current study, the consistency with which listeners with shorter intervals between sessions choose the shadowed token will be overestimated.

6. Conclusion

This study explored individual variation and consistency in listener performance in the AXB assessment of phonetic imitation. Listeners showed substantial variation in their ability to perceive imitation and that variation was relatively consistent across sessions, suggesting that the observed variation reflects differences in ability rather than random variation. This outcome provides further support to the growing body of literature indicating that individuals vary in perceptual abilities (e.g., Schertz & Clare, 2020). Furthermore, the overall conclusion about whether the shadowers imitated was the same in both sessions and the relative amount of imitation perceived among the four shadowers was the same in both sessions. These results provide evidence that the AXB assessment is a reliable tool for measuring phonetic imitation. However, due to the substantial inter-listener variation, it is recommended that future studies in which listeners respond to different subsets of trials take care to interpret their results with caution since variation between subsets could reflect differences in listener ability.

Acknowledgements

Thank you to all the participants who took part in the study, as well as the research assistants, Sophia Sedigh Afshar and Sam Turnbull, for their work in helping recruit and test the participants. Thank you also to the two anonymous reviewers for their thoughtful work in helping to improve this paper.

Competing interests

The author has no competing interests to declare.

References

Aubanel, V., & Nguyen, N. (2020). Speaking to a common tune: Between-speaker convergence in voice fundamental frequency in a joint speech production task. *PLoS ONE*, *15*(5), 1–16. DOI: https://doi.org/10.1371/journal.pone.0232209

Babel, M. (2010). Dialect divergence and convergence in New Zealand English. *Language in Society*, *39*, 437–456. DOI: https://doi.org/10.1017/S0047404510000400

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 178–189. DOI: https://doi.org/10.1016/j.wocn.2011.09.001

Babel, M., & Bulatov, D. (2011). The role of fundamental frequency in phonetic accommodation. *Language and Speech*, *55*(2), 231–248. DOI: https://doi.org/10.1177/0023830911417695

Babel, M., McAuliffe, M., & Haber, G. (2013). Can mergers-in-progress be unmerged in speech accommodation? *Frontiers in Psychology*, *4*(SEP). DOI: https://doi.org/10.3389/fpsyg.2013.00653

Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, *5*(1), 123–150. DOI: https://doi.org/10.1515/lp-2014-0006

Bannigan, K., & Watson, R. (2009). Reliability and validity in a nutshell. *Journal of Clinical Nursing*, *18*, 3237–3243. DOI: https://doi.org/10.1111/j.1365-2702.2009.02939.x

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (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

Bonin, F., De Looze, C., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., & Campbell, N. (2013). Investigating fine temporal dynamics of prosodic and lexical accommodation. *Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH, 539–543. DOI: https://doi.org/10.21437/Interspeech.2013-151

Brouwer, S., Mitterer, H., & Huettig, F. (2010). Shadowing reduced speech and alignment. *The Journal of the Acoustical Society of America*, *128*(1), EL32–EL37. DOI: https://doi.org/10.1121/1.3448022

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

Carmines, E., & Zeller, R. (1979). Reliability and Validity Assessment. SAGE Publications, Inc. DOI: https://doi.org/10.4135/9781412985642

Chandrasekaran, B., Krishnan, A., & Gandour, J. T. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain and Language*, *108*, 1–9. DOI: https://doi.org/10.1016/j.bandl.2008.02.001

Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: the perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, *76*(6), 893–910. DOI: https://doi.org/10.1037/0022-3514.76.6.893

Clopper, C. G., & Dossey, E. (2020). Phonetic convergence to Southern American English: Acoustics and perception. *The Journal of the Acoustical Society of America*, *147*(1), 671–683. DOI: https://doi.org/10.1121/10.0000555

Cohen Priva, U., Edelist, L., & Gleason, E. (2017). Converging to the baseline: Corpus evidence for convergence in speech rate to interlocutor's baseline. *The Journal of the Acoustical Society of America*, *141*(5), 2989–2996. DOI: https://doi.org/10.1121/1.4982199

Cohen Priva, U., & Sanker, C. (2018). Distinct behaviors in convergence across measures. *Proceedings of the Annual Conference of the Cognitive Science Society*, July, 1518–1523.

Cohen Priva, U., & Sanker, C. (2020). Natural leaders: some interlocutors elicit greater convergence across conversations and across characteristics. *Cognitive Science*, *44*(10). DOI: https://doi.org/10.1111/cogs.12897

Cowan, N., Elliott, E. M., Saults, S. J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. DOI: https://doi.org/10.1016/j.cogpsych.2004.12.001

Darcy, I., Mora, J. C., & Daidone, D. (2016). The role of inhibitory control in second language phonological processing. *Language Learning*, *66*(4), 741–773. DOI: https://doi.org/10.1111/lang.12161

De Krom, G. (1994). Consistency and reliability of voice quality ratings for different types of speech fragments. *Journal of Speech and Hearing Research*, *37*(5), 985–1000. DOI: https://doi.org/10.1044/jshr.3705.985

Devon, H. A., Block, M. E., Moyle-Wright, P., Ernst, D. M., Hayden, S. J., Lazzara, D. J., Savoy, S. M., & Kostas-Polston, E. (2007). A psychometric toolbox for testing validity and reliability. *Journal of Nursing Scholarship*, *39*(2), 155–164. DOI: https://doi.org/10.1111/j.1547-5069.2007.00161.x

Dias, J. W., & Rosenblum, L. D. (2016). Visibility of speech articulation enhances auditory phonetic convergence. *Attention, Perception, and Psychophysics*, *78*(1), 317–333. DOI: https://doi.org/10.3758/s13414-015-0982-6

Dias, J. W., Vazquez, T. C., & Rosenblum, L. D. (2021). Perceptual learning of phonetic convergence. *Speech Communication*, *133*, 1–8. DOI: https://doi.org/10.1016/j.specom.2021.07.004

Dufour, S., & Nguyen, N. (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology*, 4(346). DOI: https://doi.org/10.3389/fpsyg.2013.00346

Giles, H. (1973). Accent mobility: a model and some data. *Anthropological Linguistics*, *15*(2), 87–105.

Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context, and consequence. In H. Giles, N. Coupland, & J. Coupland (Eds.), Contexts of Accommodation (pp. 1–68). Cambridge University Press. DOI: https://doi.org/10.1017/CBO9780511663673.001

Giles, H., & Smith, P. M. (1979). Accommodation theory: Optimal levels of convergence. In H. Giles & R. S. Clair (Eds.), *Language and Social Psychology* (pp. 45–65). Blackwell.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. DOI: https://doi.org/10.1037/0033-295X.105.2.251

Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of Phonetics*, *32*(3), 395–421. DOI: https://doi.org/10.1016/S0095-4470(03)00016-0

Hall-Lew, L., Honeybone, P., & Kirby, J. (2021). Individuals, communities, and sound change: an introduction. *Glossa: A Journal of General Linguistics*, *6*(1), 1–17. DOI: https://doi.org/10.5334/gjgl.1630

Højen, A., & Flege, J. E. (2006). Early learners' discrimination of second-language vowels. *The Journal of the Acoustical Society of America*, *119*(5), 3072–3084. DOI: https://doi.org/10.1121/1.2184289

Jia, G., Strange, W., Wu, Y., Collado, J., & Guan, Q. (2006). Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure. *The Journal of the Acoustical Society of America*, *119*(2), 1118. DOI: https://doi.org/10.1121/1.2151806

Kim, M. (2012). Phonetic accommodation after auditory exposure to native and nonnative speech (Doctoral dissertation) (Vol. 74, Issue 2). ProQuest Dissertations Publishing.

Kim, D., & Clayards, M. (2019). Individual differences in the relation between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769–786. DOI: https://doi.org/10.1121/1.4969741

Kim, D., Clayards, M., & Goad, H. (2017). Individual differences in second language speech perception across tasks and contrasts: The case of English vowel contrasts by Korean learners. *Linguistics Vanguard*, *3*(1), 20160025. DOI: https://doi.org/10.1515/lingvan-2016-0025

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. *Journal of Phonetics*, *59*, 40–57. DOI: https://doi.org/10.1016/j.wocn.2016.08.006

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). DOI: https://doi.org/10.18637/jss.v082.i13

Kwon, H. (2021). A non-contrastive cue in spontaneous imitation: Comparing mono- and bilingual imitators. *Journal of Phonetics*, *88*, 101083. DOI: https://doi.org/10.1016/j.wocn.2021.101083

Larraza, S., & Best, C. T. (2018). Differences in phonetic-to-lexical perceptual mapping of L1 and L2 regional accents. *Bilingualism*, *21*(4), 805–825. DOI: https://doi.org/10.1017/S1366728917000323

Lev-Ari, S. (2018). Social network size can influence linguistic malleability and the propagation of linguistic change. *Cognition*, *176*, 31–39. DOI: https://doi.org/10.1016/j.cognition.2018.03.003

Lev-Ari, S., & Peperkamp, S. (2014). An experimental study of the role of social factors in language change: The case of loanword adaptations. *Laboratory Phonology*, *5*(3). DOI: https://doi.org/10.1515/lp-2014-0013

Lewandowski, E. M., & Nygaard, L. C. (2018). Vocal alignment to native and non-native speakers of English. *The Journal of the Acoustical Society of America*, *144*(2), 620–633. DOI: https://doi.org/10.1121/1.5038567

Lewandowski, N., & Jilka, M. (2019). Phonetic convergence, language talent, personality and attention. *Frontiers in Communication*, *4*(18), 1–19. DOI: https://doi.org/10.3389/fcomm.2019.00018

Lin, Y., Yao, Y., & Luo, J. (2021). Phonetic accommodation of tone: Reversing a tone merger-in-progress via imitation. *Journal of Phonetics*, *87*, 101060. DOI: https://doi.org/10.1016/j.wocn.2021.101060

MacLeod, B. (2014). Investigating the effects of salience and regional dialect on phonetic convergence in Spanish. In M.-H. Côté & E. Mathieu (Eds.), Variation within and across Romance Languages: Selected papers from the 41st Linguistic Symposium on Romance Languages (LSRL), Ottawa, 5–7 May 2011 (pp. 351–378). John Benjamins. DOI: https://doi.org/10.1075/cilt.333.24mac

MacLeod, B., & Di Lonardo Burr, S. M. (2022). Phonetic imitation of the acoustic realization of stress in Spanish: Production and perception. *Journal of Phonetics*, *92*, 101139. DOI: https://doi.org/10.1016/j.wocn.2022.101139

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. In *Behavior Research Methods* (Vol. 44, Issue 2, pp. 314–324). DOI: https://doi.org/10.3758/s13428-011-0168-7

Mayr, R., & Escudero, P. (2010). Explaining individual variation in L2 perception: Rounded vowels in English learners of German. *Bilingualism*, *13*(3), 279–297. DOI: https://doi.org/10.1017/S1366728909990022

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2013). Is speech alignment to talkers or tasks? *Attention, Perception & Psychophysics*, *75*(8), 1817–1826. DOI: https://doi.org/10.3758/s13414-013-0517-y

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

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. DOI: https://doi.org/10.1016/j.wocn.2010.12.007

Nielsen, K., & Scarborough, R. (2019). Perceptual target of phonetic accommodation: A pattern within a speaker's phonetic system or the raw acoustic signal? *Proceedings of the 19th International Congress of Phonetic Sciences* (ICPhS 2019) (Melbourne).

Nye, P. W., & Fowler, C. A. (2003). Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English. *Journal of Phonetics*, *31*, 63–79. DOI: https://doi.org/10.1016/S0095-4470(02)00072-4

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*, 2382–2393. DOI: https://doi.org/10.1121/1.2178720

Pardo, J. S. (2012). Reflections on phonetic convergence: Speech perception does not mirror speech production. *Language and Linguistics Compass*, *6*(12), 753–767. DOI: https://doi.org/10.1002/lnc3.367

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4*(AUG), 1–5. DOI: https://doi.org/10.3389/fpsyg.2013.00559

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197. DOI: https://doi.org/10.1016/j.wocn.2011.10.001

Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. Journal of Memory and Language, *69*(3), 183–195. DOI: https://doi.org/10.1016/j.jml.2013.06.002

Pardo, J. S., Urmanche, A., Wilman, S., & Wiener, J. (2017). Phonetic convergence across multiple measures and model talkers. *Attention, Perception, and Psychophysics*, *79*(2), 637–659. DOI: https://doi.org/10.3758/s13414-016-1226-0

Pardo, J. S., Urmanche, A., Wilman, S., Wiener, J., Mason, N., Francis, K., & Ward, M. (2018). A comparison of phonetic convergence in conversational interaction and speech shadowing. *Journal of Phonetics*, *69*, 1–11. DOI: https://doi.org/10.1016/j.wocn.2018.04.001

Phillips, S., & Clopper, C. G. (2011). Perceived imitation of regional dialects. *Proceedings of Meetings on Acoustics*, *12*. DOI: https://doi.org/10.1121/1.4704668

Polka, L. (1995). Linguistic influences in adult perception of non-native vowel contrasts. *The Journal of the Acoustical Society of America*, *97*(2), 1286–1296. DOI: https://doi.org/10.1121/1.412170

Ross, J. P., Lilley, K. D., Clopper, C. G., Pardo, J. S., & Levi, S. V. (2021). Effects of dialect-specific features and familiarity on cross-dialect phonetic convergence. *Journal of Phonetics*, *86*, 101041. DOI: https://doi.org/10.1016/j.wocn.2021.101041

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. DOI: https://doi.org/10.1016/j.wocn.2015.07.003

Schertz, J., & Clare, E. J. (2020). Phonetic cue weighting in perception and production. *WIREs Cognitive Science*, *11*(2), e1521. DOI: https://doi.org/10.1002/wcs.1521

Schertz, J., & Johnson, E. K. (2022). Voice onset time imitation in teens versus adults. *Journal of Speech, Language, and Hearing Research*, *65*(5), 1839–1850. DOI: https://doi.org/10.1044/2022_JSLHR-21-00460

Schweitzer, A., & Walsh, M. (2016). Exemplar dynamics in phonetic convergence of speech rate. *Proceedings of the Annual Conference of the International Speech Communication Association*, INTERSPEECH, 08-12-Sept, 2100–2104. DOI: https://doi.org/10.21437/Interspeech.2016-373

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429. DOI: https://doi.org/10.3758/BF03194890

Smith, R. (2015). Perception of speaker-specific phonetic detail. In S. Fuchs, D. Pape, C. Petrone, & P. Perrier (Eds.), Individual Differences in Speech Production and Perception (pp. 11–38). Frankfurt am Main.

Sonderegger, M., Wagner, M., & Torreira, F. (2018). Quantitative Methods for Linguistic Data (1.0).

Stewart, M. E., & Ota, M. (2008). Lexical effects on speech perception in individuals with ''autistic" traits. *Cognition*, *109*(1), 157–162. DOI: https://doi.org/10.1016/j.cognition.2008.07.010

Street, R. L. (1982). Evaluation of noncontent speech accommodation. *Language & Communication*, *2*(1), 13–31. DOI: https://doi.org/10.1016/0271-5309(82)90032-5

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. DOI: https://doi.org/10.1080/17470218.2013.850521

Wade, L., Lai, W., & Tamminga, M. (2020). The reliability of individual differences in VOT imitation. *Language and Speech*, 1–18. DOI: https://doi.org/10.1177/0023830920947769

Wagner, M. A., Broersma, M., McQueen, J. M., Dhaene, S., & Lemhöfer, K. (2021). Phonetic convergence to non-native speech: Acoustic and perceptual evidence. *Journal of Phonetics*, *88*, 101076. DOI: https://doi.org/10.1016/j.wocn.2021.101076

Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, *6*(Article 546), 1–18. DOI: https://doi.org/10.3389/fpsyg.2015.00546

Werker, J. F., & Logan, J. S. (1985). Cross-language evidence for three factors in speech perception. *Perception and Psychophysics*, 37, 35–44. DOI: https://doi.org/10.3758/BF03207136

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag. DOI: https://doi.org/10.1007/978-3-319-24277-4

Yu, A. C. L., Abrego-Collier, C., & Sonderegger, M. (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "Autistic" traits. *PLoS ONE*, *8*(9), e74746. DOI: https://doi.org/10.1371/journal.pone.0074746

Yu, A. C. L., & Zellou, G. (2019). Individual differences in language processing: Phonology. *Annual Review of Linguistics*, *5*, 131–150. DOI: https://doi.org/10.1146/annurev-linguistics-011516-033815

Zellou, G., Cohn, M., & Ferenc Segedin, B. (2020). Age- and gender-related differences in speech alignment toward humans and voice-AI. *Frontiers in Communication*, *5*. DOI: https://doi.org/10.3389/fcomm.2020.600361

Zellou, G., Dahan, D., & Embick, D. (2017). Imitation of coarticulatory vowel nasality across words and time. *Language, Cognition and Neuroscience*, *32*(6), 776–791. DOI: https://doi.org/10.1080/23273798.2016.1275710

Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *The Journal of the Acoustical Society of America*, *140*(5), 3560–3575. DOI: https://doi.org/10.1121/1.4966232