Title

Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication

Author names and affiliations

Gwen Brekelmans (corresponding author)

Department of Biological and Experimental Psychology, Queen Mary University of London, Fogg Building, Mile End Road, London E1 4NS, United Kingdom

g.brekelmans@qmul.ac.uk

Nadine Lavan

Department of Biological and Experimental Psychology, Queen Mary University of London, Fogg Building, Mile End Road, London E1 4NS, United Kingdom

n.lavan@qmul.ac.uk

Haruka Saito

Département de linguistique, Université du Québec à Montréal, 320, rue Sainte-Catherine Est, Montréal, QC, Canada, H2X 1L7

saito.haruka@courrier.uqam.ca

Meghan Clayards

Department of Linguistics, 1085 Ave Dr. Penfield, Montréal, QC H3A 1A7, Canada

School of Communication Sciences & Disorders, 2001 McGill College, 8th floor, Montréal, QC H3A 1G1, Canada

meghan.clayards@mcgill.ca

Elizabeth Wonnacott

Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom

elizabeth.wonnacott@education.ox.ac.uk

[NOTE: THIS REGISTERED REPORT HAS BEEN PEER-REVIEWED AND HAS RECEIVED IN PRINCIPLE ACCEPTANCE AT STAGE 1, AND STAGE 2 HAS BEEN ACCEPTED FOR PUBLCATION AT IN THE JOURNAL OF MEMORY AND LANGUAGE.]

Author Note

The registration for this report, as well as all stimuli, data, and analyses scripts, can be found on the Open Science Framework (OSF) page of this project: https://osf.io/4hkqb/

Abstract

Acquiring non-native speech contrasts can be difficult. A seminal study by Logan, Lively and Pisoni (1991) established the effectiveness of *phonetic training* for improving non-native speech perception: Japanese learners of English were trained to perceive /r/-/l/ using minimal pairs over 15 training sessions. A pre/post-test design established learning and generalisation. In a follow up study, Lively, Logan and Pisoni (1993) presented further evidence which suggested that talker variability in training stimuli was crucial in leading to greater generalisation.

These findings have been very influential and “high variability phonetic training” is now a standard methodology in the field. However, while the general benefit of phonetic training is well replicated, the evidence for an advantage of high over lower variability training remains mixed. In a large-scale replication of the original studies using updated statistical analyses we test whether learners generalise more after phonetic training using multiple talkers over a single talker. We find that listeners learn in both multiple and single talker conditions. However, in training, we find no difference in how well listeners learn for high vs low variability training. When comparing generalisation to novel talkers after training in relation to pre-training accuracy, we find ambiguous evidence for a high-variability benefit over low-variability training: This means that if a high-variability benefit exists, the effect is much smaller than originally thought, such that it cannot be detected in our sample of 166 listeners.

Introduction

Learning to perceive non-native speech sounds can be difficult. While there is generally learning over time (e.g. McKain et al., 1981), problems can persist even after months or years of learning in an immersion environment (Flege & MacKay, 2004). This leads to the question of whether targeted phonetic training can be useful to support learning, thus enhancing speech comprehension. During phonetic training, listeners are trained to

discriminate non-native speech contrasts via minimal pairs (e.g. ’rock’ and ‘lock’). The foundations of this kind of phonetic training were established in the 1970s and 1980s. In an early influential study, Strange and Dittmann (1984) attempted to implement minimal-pair phonetic training using tokens from a continuum of synthesised speech to train Japanese learners of English on the /l/-/r/ contrast. This contrast is notoriously difficult for these learners, as Japanese only has one phoneme (/r/) where English uses two (/l/-/r/). Strange and Dittmann (1984) showed that learners significantly improved in their /l/-/r/ discrimination and identification abilities after such training, but that improvements, crucially, did not generalise to natural speech. However, generalisation to natural speech was shown in a later study (Jamieson & Morosan, 1986) that used a training paradigm that gradually introduced more varied synthesised stimuli.

A turning point in the phonetic training literature occurred in the early 1990s, when researchers began using natural speech in phonetic training. Logan et al. (1991) again focussed on Japanese learners of English acquiring /l/-/r/, but in contrast to previous training studies used natural speech where the contrasts occurred in multiple contexts and were spoken by multiple speakers. Training took place over fifteen sessions and consisted of an identification task with trial-by-trial feedback on their performance. Learning was assessed using a pre-test/post-test design, where participants were given the same battery of tests before and after training so that improvement could be measured. Afterwards, a generalisation task was administered, which tested their perception of this speech contrast using items and talkers that did not occur in training. The authors found that participants improved on trained stimuli as well as untrained stimuli using both novel voices as well as novel items. In a follow up study, the authors examine the effect of different types of variability in the input stimuli (Lively et al., 1993). They report that participants’ ability to generalise to novel stimuli produced by a new talker was specifically dependent on exposure to training stimuli spoken by multiple talkers rather than a single talker, while the manipulation of phonetic environment did not influence improvement. Theoretically (and

intuitively), it makes sense to see such an advantage of talker variability input on generalisation tasks. Encountering variability helps the listener to recognise which acoustic cues are irrelevant to accurate discrimination, and thus to focus on those cues that are key to distinguish the phonemes they are being trained on. The logic is that any two speakers will likely sound a little different. Thus, if you only ever heard L2 phonemes produced by one talker, you might think that aspects of that speaker's pronunciation are relevant to distinguishing the phonemes, when they are actually idiosyncrasies of the speaker. This could make it difficult to adjust to novel speakers with different talker-specific idiosyncrasies from the talker you have heard. In contrast, when you are exposed to multiple speakers who all pronounce the phonemes slightly differently, you can learn which cues are variable and thus irrelevant, and instead focus on those cues that are diagnostic.

Since this seminal work, *high variability phonetic training* (HVPT) has become a standard paradigm in the field. A number of studies from the same research group have produced additional evidence to show that the paradigm is effective: Lively et al. (1994) report that improvements in /l/-/r/ perception were fully retained at a three-month follow-up, and although a six-month follow-up showed some decreased performance, learners remained well above their initial pre-test level. Bradlow et al. (1997) show perceptual training transferred to improvements in production on the trained /l/-/r/ contrast, a finding that is extended by Bradlow et al. (1999), who show long-term retention of the /l/-/r/ contrast in both perception and production at a three-month follow up. In later years these findings have been replicated and extended to other consonant contrasts (e.g. Fuhrmeister & Meyers, 2017), vowel contrasts (e.g. Nishi & Kewley-Port, 2007), lexical tones (e.g. Sadakata & McQueen, 2014), and larger discourse contexts (Huensch, 2016), and different populations (e.g. child learners; Heeren & Schouten, 2010). In sum, there is therefore substantial evidence that HVPT can be effective, and it is frequent used to improve the acquisition of non-native contrasts (see e.g. Sakai & Moorman 2018 for a meta-analysis). An extended overview of all phonetic training

studies that have cited one or both of the original studies is provided in Appendix A ('Literature search').

The impact of these seminal studies is further illustrated by their citation counts: as of 12 December 2020, Web of Science currently shows 351 citations for Logan et al. (1991) and 355 for Lively et al. (1993), while Google Scholar citations are at 907 for Logan et al. (1991), and at 777 for Lively et al. (1993). Moreover, the studies have impacted theoretical, empirical and applied aspects (e.g. see Thomson, 2018) of non-native speech perception, language learning, and beyond. However, while all studies reviewed above use high variability input, introduced by including multiple talkers during training, the majority have *not* returned to the question of whether variability in the input is indeed important for learning. HVPT is in all cases assumed to be the most effective type of training but potential differences between high and low variability during training are not tested for. In fact, the original studies by Logan et al. (1991) and Lively et al. (1993) are cited in the majority of the later papers to motivate the choice of using high variability in their phonetic training materials.

The goal of the current paper is to replicate the key experiments from Logan et al. (1991) and Lively et al. (1993) and test the hypothesis that using high variability input (specifically, input with multiple talkers) leads to more generalisation in phonetic learning than when low variability input is used. In the remainder of this introduction, we first discuss the key experiments contrasting high and low variability in Logan et al. (1991) and Lively et al. (1993). We then turn to studies that have consequently investigated effects of talker variability in the literature, starting with a review of studies which – as in the original studies and this replication – focussed on phonetic training and then turning to studies in related areas of voice identification, dialect identification and vocabulary learning. Finally, as a counterpoint to the literature on high-variability benefits, we review intriguing evidence from an adjacent literature on spoken language processing in which high variability training has in fact been shown to be at times detrimental.

Logan et al., (1991) and Lively et al., (1993): Contrasting multiple talker and single talker training

Logan et al. (1991) and Lively et al. (1993) are the two seminal studies reporting to find evidence that high variability training improved non-native phoneme perception, and these studies thus form the basis of our replication. Across the two papers there are three different experiments, run with different groups of participants. The first experiment (Logan et al., 1991), used training stimuli produced by multiple (five) talkers with multiple phonetic contrasts. The second experiment, (Experiment 1 in Lively et al., 1993) used training stimuli produced by multiple (five) talkers with more limited variation in phonetic contrasts, and the third experiment (Experiment 2 in Lively et al., 1993) used training stimuli produced by a single talker while using multiple phonetic contrasts. The strongest claim regarding the beneficial effects of high input variability concerned the contrast between the first and third of these, i.e. the multiple talker input (high variability) versus single talker input (low variability), and we will focus on these in this literature review and in our replication.

Experimental design

In both experiments, Japanese learners of English were trained on the /l/-/r/ contrast through fifteen phonetic training sessions. A single training session consisted of a two-alternative forced-choice (2AFC) identification task in which learners were played a word that contains either an /l/ or an /r/, and they were asked to identify which of the two they think they heard. The two experiments crucially differ in their training input: Logan et al. (1991) used multiple talker training where listeners hear five different voices, while Lively et al. (1993) used a single talker training throughout all fifteen training sessions. Beyond this, both studies used the same pre/post paradigm. Both experiments started with a 2AFC /l/-/r/ identification task at pre-test, using untrained items taken from Strange and Dittmann (1984). This test was repeated after training and is referred to as the post-test. This post-test again used untrained items and an untrained talker, though note that the talker is the same as the one used at pre-test. The post-test was followed by two further tasks to measure generalisation:

generalisation 1 which tests novel items with a novel talker, and generalisation 2 which tests a different set of novel items but with a talker that was used in training. Using this paradigm, the authors hoped to unpack whether any improvement in perception of the non-native contrast that was seen in training might also generalise, and thus extend to items and voices that learners had not heard before.

Findings

Logan et al. (1991) report that the multiple talker phonetic training paradigm was successful at improving the learners' abilities to identify /l/ and /r/ from pre- to post-test. Phonetic context was found to affect performance, as were specific talkers. Importantly, performance in the generalisation test was greater for a familiar talker from training than for a talker who was unfamiliar. The authors interpret this as evidence that learners encode and store talker-specific information. They more generally speculate that for training to be effective and lead to generalisation to a range of talkers, the training input needs to contain sufficient variation. Lively et al. (1993) describe that in the single talker version of the phonetic training paradigm, learners did improve from pre- to post-test, but that, in generalisation, performance with a novel talker was lower than performance with a familiar talker. Generalisation performance is referred to as being "mediocre" as performance was similar to the first week of training, thus suggesting that improvements during training were specific to the stimuli rather than more robustly generalisable. Referring back to the findings of Logan et al. (1991), they conclude that talker variability in the training input must be important for attaining such robust generalisation.

When examining the results, however, it becomes apparent that the claim that talker variability is essential for generalisation is not well-supported. Logan et al. (1991) saw higher raw scores on tasks with untrained talkers after training than the participants in the single talker training condition from Lively et al. (1993) (i.e. 86% vs 80% for post-test and 80% vs 76% for generalisation 1). However, one of these tests - generalisation 1, the task they consider most critical for generalisation - was only administered at post-test and thus pre- to

post-test improvement is not measured. It is thus possible that higher scores in this test are not the result of the different levels of variability in training, but due to chance differences between the group of participants tested by Logan et al. (1991) and those tested by Lively et al. (1993). In fact, the available pre-test data suggests that this may indeed be the case, as performance at pre-test – prior to any training – in the multiple talker study is 78% while that in the single talker study is 69%. However, since this pre-test task does not use the same items as the generalisation task no strong conclusions can be drawn from this.

Two other aspects of the two seminal studies make it difficult to fully interpret the reported effects of talker variability on the learning of non-native phonetic contrasts. First, only six participants were tested per experiment in both Logan et al. (1991) and Lively et al. (1993). Furthermore, Logan et al. (1991) only administered generalisation tasks on three of those six participants. Second, the reported high variability benefit for generalisation was only described and, critically, was never tested statistically as the two experiments were analysed separately across two papers. Given these weaknesses in the experimental design, we maintain that - notwithstanding the wide-reaching impact of this claim by the authors - it is not possible to draw firm conclusions from these two studies about whether there is truly a benefit for multiple talkers over single talker input in phonetic training.

Literature Review: Phonetic training studies contrasting high and low variability

We will now turn to consider whether any subsequent phonetic training studies have demonstrated a benefit for multi-talker, high variability input over single-talker low variability input. To identify all relevant studies, we conducted a systematic literature review for which we searched Web of Science for the studies that had cited the original two papers, exporting 692 entries on 15 December 2020. Duplicates were removed, leaving 527 entries that cited either of the two original studies. For these 527 entries, abstracts and keywords were inspected by the first author to determine whether studies used phonetic training and whether they contrasted input variability. Where abstracts suggested phonetic training might be used but exact procedure was unclear, the full articles were inspected to confirm. Of

these 527 possible entries, 17 contrast high and low talker variability. Of these, two were review papers or chapters where the talker variability section did not present novel experimental results, and one was a published conference proceedings paper that was later extended into a full journal article which was also on the list. This left 12 published papers which will be discussed below. Note that in addition to these 12 papers, two of the current authors (GB and EW) have been involved in further research directly contrasting high and low variability in phonetic training. These studies did not come up through the current search as the studies have so far been disseminated via conference proceedings and/or are under review (pre-prints available), however they have also been added to the discussion below due to their direct relevance. The full overview of relevant studies can be found in Appendix A.

Adult learners

Our literature search showed that for phonetic training of non-native segmental contrasts, very few further studies with adult participants have directly contrasted multiple talker (high variability) and single talker (low variability) input; all other studies only included multiple talker input. A first study to contrast single versus multiple talker input was a training study by Sadakata and McQueen (2013) in which Dutch adult participants were trained to acquire a Japanese geminate consonant contrast in five training sessions. Training input was either high variability – with both multiple talkers and more varied items – or low variability – with a single talker and less varied items. Participants showed better learning as well as generalisation in an identification task after high variability training than low variability training, although their discrimination performance improved regardless of training variability. Similarly, Wong (2012) trained Cantonese learners on English vowel contrasts for ten sessions of high variability (multiple talkers, multiple phonetic contexts) or low variability (single talker, one phonetic context) input. Both groups improved after training, although those with high variability input outperformed those with low variability input and showed

greater generalisation on identification tasks. This result was then replicated with learners of high and low proficiency levels in Wong (2014).

There are two additional studies training segmental perception more broadly that have varied single versus multiple talkers in the training input. The first is Hardison (2003), who trained Korean learners on English /r/-/l/. The study compared audio-visual or audio-only perceptual training, examining if there was an audio-visual benefit both when training with multiple talkers and a single talker. Performance in generalisation showed a marginal benefit for multiple talker input in the audio-visual but not audio-only training condition. Overall performance on the generalisation task with both novel talkers and items was lower than on the task using a familiar talker but novel items. A second study that uses single versus multiple talker input in perceptual training is Brosseau-Lapré et al., (2013), who examined the learning of a novel vowel contrast under different variability conditions. Specifically, English learners of a French vowel contrast were given two training sessions, where talker variability was either high (multiple talkers) or low (a single talker). Analyses looked at pre- to post-change in participants identification functions and found statistically significant training effects only in conditions with multiple talkers. However, no direct comparison for single versus multiple talker input was made across training conditions.

Turning to speech production, Kartushina and Martin (2019) used a phonetic training task to see whether knowledge gained through perception-based training generalised to production. They showed that two sessions of listening to either high (multiple talker) or low (single talker) variability stimuli both improved production accuracy in adult Spanish learners of a French vowel contrast. However, only high variability training resulted in generalisation when participants were asked to repeat items spoken in a novel voice. Additionally, the high variability training led to more stable production of the trained vowels than low variability training.

The role of variability has also been investigated in phonetic training of lexical tone perception. Perrachione et al. (2011) trained English learners on Mandarin tones in eight

training sessions. They found that when there was trial-by-trial variability with talkers changing at every trial, high variability multi-talker training input was beneficial specifically for learners who had stronger perceptual abilities (as measured on a pitch-contour perception test), while high variability multi-talker input had a detrimental effect on learners with weaker perceptual abilities. Sadakata and McQueen (2014) came to a similar conclusion in their five-session training of Mandarin tones: Dutch learners who had a low perceptual aptitude were hindered by increased talker variability during training, while those with a high aptitude benefitted from it. These results are consistent with the greater difficulties that non-native speakers have in processing multiple talker than single talker input (reviewed in the section on language processing below), and suggest this may offset potential benefits of variability in less able participants. However, a two-session study training Mandarin learners on Cantonese tones found no difference in high or low talker variability on tone identification, regardless of participants' perceptual ability (Zhang et al., 2018). The only effect of variability seen in this study was a benefit of high variability on learning to produce two of the six tones, but again no link to perceptual aptitude was found. Similarly, an eight-session tone training study for English learners of Mandarin tones by Dong et al. (2019) did not find either an overall benefit of training with high variability (multiple talker) materials, nor an interaction with individual aptitude. By using Bayes Factors, Dong et al. (2019) in fact demonstrate substantial evidence for the null hypothesis for the prediction that there would be a high variability benefit. For the predicted interaction between variability and aptitude, the evidence was ambiguous. Two further tone training studies by Wiener et al. (2020) and Deng et al. (2018) contrasted high and low variability input and did not find a significant difference due to talker variability (they did not investigate the interaction with individual perceptual difference). Deng et al. (2018) found no difference in generalisation performance between high or low talker variability conditions. Wiener et al. (2020) similarly found no effect of talker variability overall, but found there was an interaction with the use of explicit instruction and tone: Multiple talker input only proved beneficial on one of the four trained tones if there was additional explicit instruction.

Child learners

There is also a growing literature on phonetic training in second language learning in children. However, similar to the literature on adult learners, only very few studies have directly compared high and low variability training in child learners. An experiment reported in Evans and Martín-Alvarez (2016) and further expanded in Evans et al. (2017) trained Spanish children aged 9-12 on an English vowel contrast through means of either high variability (four talkers) or low variability (single talker) perceptual training input using a whole-picture identification task. The aim of this study was to improve perception on a discrimination task as well as to investigate any transfer to production. Only children trained on a single talker improved in production, while multiple talker input played a role in improvement of vowel perception. Specifically, this study found no overall multiple talker benefit for generalisation to perception of novel voices alone, however a multiple talker benefit was seen specifically for tests extending to novel items. This suggests that the high variability benefit might only be seen when stimuli are further removed from the training set, requiring more generalisation. A second study (Giannakopoulou et al., 2017) used the same training stimuli and methods as Evans and Martín-Alvarez (2016), with additional training sessions, and included children and adults. Unlike Evans and Martín-Alvarez (2016), this study did *not* find a multiple talker benefit in the discrimination test. In fact, children showed an unexpected single-talker benefit in the perceptual discrimination task that even held for generalisation across novel speakers and items, while for adults there were no differences between the conditions. The authors were cautious in interpreting these results: For adults, they note that the lack of a high variability, multiple talker benefit could be due to the fact that performance was near ceiling at post-test. For children, they acknowledge that the unexpected single talker benefit might be due to accidental differences between the groups at pre-test.

Finally, Brekelmans et al. (2020) also investigate the role of high versus low talker variability in 7-year-olds and 11-year-old Dutch children learning English vowels. They found that both

age groups improved as a result of eight training sessions, but 7-year-olds did not show any generalisation abilities regardless of which input variability condition they were in. The 11-year-olds did show generalisation abilities, but crucially did not show a benefit of high variability (four talkers) over low variability (one talker) training input. Using Bayes Factors, Brekelmans and colleagues showed that there was evidence for the null for a multiple talker benefit in some of the tasks.

In summary, the studies that contrast high versus low talker variability in phonetic training are relatively few in number and their results are not consistent. For adult learners, some studies report high-variability training benefits, although for the majority of studies reviewed the evidence for a clear benefit is either not compelling or entirely absent. For child learners, there appears to be less evidence for high variability training benefits, with many surprising findings, such as reports of low variability training benefits.

Tests of talker variability in other areas of speech-based learning

Variability effects have also been reported in other domains beyond phonetic training. In reflection of this wide-reaching impact of the original studies reporting high variability training benefits, we will briefly review relevant literature which has looked at benefits of multiple talkers in related areas of research outside of phonetic training. These studies were identified within the same literature search detailed above. Through this process, 24 studies in which talker-related variability played a role were identified. One of these was a published conference proceedings paper that was later extended into a full journal article which was also on the list; this left 23 papers which will be discussed below. As before, we will summarise these studies briefly below.

Our literature search found studies in the areas of voice and dialect identification. Lavan et al. (2019) report hearing varied speaking styles while learning new voice identities was beneficial compared to hearing a single speaking style. This high variability benefit was most pronounced when generalisation was required. Note, however, that listeners trained on one speaking style showed a low variability benefit when test items were similar (in speaking

style) to the trained stimuli. Note that this study makes reference to the face perception literature, where high variability advantages have also been described in a largely independent set of studies looking at face identity learning (Baker et al., 2015; Ritchie & Burton, 2017; Murphy et al., 2015). For dialect identification, Clopper and Pisoni (2004) found a benefit of hearing multiple voices during training compared to hearing a single voice. Crucially, this high variability multiple talker benefit was seen in generalisation, when participants were asked to categorise novel talkers into dialect regions.

Another set of studies look at talker variability in the perception of accented speech. Bradlow and Bent (2008) investigated listeners' adaptability to foreign-accented English that varied in intelligibility, looking at both talker-dependent adaptation and talker-independent adaptation. Talker-dependent adaptation, where listeners had to adapt to different accents of specific talkers, proved highest in the single-talker condition. For talker-independent adaptation, where listeners adapted to a broader accent spoken by a group of talkers, hearing multiple talkers was beneficial for generalisation to a novel talker. Similarly, Gao et al. (2013) investigated the benefit of single versus multiple talker exposure in adapting to foreign-accented English. They found that for learners who were more advanced, multiple talker input was useful for performance in tasks with both a familiar and a novel talker, while for learners with lower proficiency single talker input resulted in better performance overall [this finding is reminiscent of the results of Perrachione et al. (2011) and Sadakata and McQueen (2014) discussed above]. Furthermore, a benefit of exposure to multiple talkers over single talkers was found in accented word recognition in 18-month-old infants (Potter & Saffran, 2017).

Variability has also been investigated in the area of cochlear implant and hearing aid research. Barcroft et al. (2011) report that hearing aid users trained on single or multiple talker speech showed similar improvements during speech perception. Each condition performed best on the version of the test that had the same number of speakers as they had been trained on (i.e. single talker training resulted in better single talker test performance).

Stacey and Summerfield (2007) explored the role of talker variability in the ability to perceive noise-vocoded speech (i.e. speech distorted in such a way as to simulate the effect of a cochlear implant). Training with multiple talkers was found to be beneficial for generalisation to novel talkers, but only when the distortion of the input was sufficiently low.

Another area in which the effects of high versus low talker variability have been explored is research looking more broadly at how languages are influenced by the populations who use them. In an artificial language study, Lev-Ari (2018) found that participants who had larger social networks (and are thus regularly exposed to more talker variation) performed better on a speech in noise task. This result was extended in Lev-Ari and Sebanz (2020), where participants who interacted with multiple compared with a single communication partner were better understood and performed better in a memory task based on descriptions they gave to a single or multiple partners. These results were broadly interpreted as demonstrating that increased variability is beneficial for communication skills.

There is also a substantial body of evidence indicating that input variability plays a role in vocabulary learning: A number of studies have shown that adults show better recall of L2 vocabulary after multiple talker input (Barcroft & Sommers, 2005; Sommers & Barcroft, 2011). This variability benefit seems to only be found when the variability is phonetically relevant, e.g. when there was variation in speaking rate which is phonetically relevant for English speakers, while variability in a cue that was not phonetically relevant, e.g. f0 for English speakers, did not result in any effect on word learning (Sommers & Barcroft, 2006, 2007; Barcroft & Sommers, 2014). This aligns with an explanation in which encountering variability is beneficial because it provides evidence about which cues are relevant: for cues that the learner expects to be irrelevant given their language background, added variability will not make a difference. Note, however, that variability benefits for vocabulary learning may be constrained: Sinkevičiūtė et al. (2019) found that while adults showed the predicted variability benefit in recall in L2 vocabulary learning, 7-8 year-olds and 10-11 year-olds did *not* show a variability effect at all (with Bayes Factors showing evidence for the null for the

former group). This result is again reminiscent of the findings reported by Perrachione et al., (2011) and Sadakata and McQueen (2014), and again can be interpreted in terms of a trade-off between the more demanding nature of processing multiple talker input (see the section on talker variability in language processing below) and the potential benefits this variable input might have for generalisation. In this context, properties of the learner (such as their age) might tip the balance towards variability becoming a burden rather than a benefit.

Other studies using artificial languages have investigated whether multiple talkers could aid both word segmentation and morphological learning in an artificial speech stream. Atkinson et al. (2015) found no benefit for talker variability in either word segmentation or morphological learning, however Estes and Lew-Williams (2015) found that infants were able to segment words from the speech stream, and to generalise these to words spoken by a novel talker, after multiple talker input but not after single talker input.

Finally, a benefit of multiple talkers has also been reported in the related area of infant language processing. Rost and McMurray (2009) proposed that a lack of input variability might play a role in the difficulty infants showed in learning minimal pairs in word learning experiments. They tested this by teaching infants novel words while contrasting single-talker and matched multiple-talker input. Infants in the first group failed to discriminate between the minimal pairs at test whilst those who received matched multiple talker input were able to correctly map the minimal pairs to different referents. On the other hand, Quam et al. (2021) expanded on this by testing whether single versus multiple talker input had an effect on native and non-native consonant contrast perception in 7-month olds. No differences between the number of talkers in habituation was seen, neither for native nor for non-native contrasts.

In addition to human perception studies, computational models have been developed that provide an account of how with single-talker input, irrelevant talker-specific acoustic features become associated with the concepts being learnt, to the detriment of relevant phonetic features, which does not happen to the same extent for multiple talker input during infant

word learning (Apfelbaum & McMurray, 2011; Rost & McMurray, 2010). This erroneous association between talker-specific features and the concepts being learnt can, however, be avoided by deliberately introducing increased acoustic variation in single-talker input (Galle et al., 2015). This computational approach is compatible with a broader class of models of linguistic learning in which linguistically relevant and irrelevant cues compete and "generalisation" occurs via a discriminative process which dissociates the irrelevant features (e.g. Ramscar et al., 2010).

In summary, high variability benefits have been reported in training studies looking at accent and dialect identification, accented word recognition, perception of speech distorted by cochlear implants, infant word segmentation and vocabulary learning in adults and infants. For the last of these, computational models have been presented which capture this hypothesized talker-variability benefit. However, as with the phonetic training literature, there is inconsistency, with some papers reporting evidence for the variability benefit, and some reporting null findings.

Tests of talker variability in language processing

The review above has focused on evidence that multiple talker input in training leads to better generalisation at post-test. However, some of the studies discussed also looked at performance during the training stage, generally finding stronger performance for single talker than multi-talker input (c.f. Brekelmans et al., 2020; Dong et al., 2019; Giannakopoulou et al., 2017; Sinkevičiūtė et al., 2019; Sadakata & McQueen, 2014; Perrachione et al., 2011). We saw above that some researchers have suggested that there may be a trade-off between the benefits of high variability for generalisation and the cost of processing multiple talker input during phonetic/vocabulary training (Perrachione et al., 2011; Sinkevičiūtė et al., 2019; Sadakata & McQueen, 2014).

Such a "cost" of high variability has also been reported in the field of language processing, where there is evidence that speech processing is more difficult with multiple talkers compared to a single talker, even in a listeners' native language. Studies discussed in this

section were identified in the same literature search described above. Talker variability that varies per trial has been shown to be detrimental in spoken word recognition (Mullenix et al., 1989), when compared to listeners who received just single talker input. Heald & Nusbaum (2014) found in a speeded word-monitoring task that having multiple talker input was detrimental compared to a single talker input in both audio-only and audio-visual conditions, the latter indicating that having the ability to see a talker's face did not alleviate any added cost of processing multiple talkers. Similar detrimental results were seen in word recall tasks, although multiple talker input was only detrimental for the first part of word lists and the difference between single or multiple talker input disappeared for later parts of the lists (Martin et al., 1989). The authors suggested processing input from multiple talkers might take up more working memory capacity, as the listener has to adjust to a novel speaker on every trial, however this effort is reduced with increased exposure.

Multi-talker input may also cause processing difficulties for non-native listeners. Lee et al. (2009) investigated the role of talker variability in the identification of acoustically manipulated Mandarin tones by native and non-native listeners. Tones produced by a single talker were identified more accurately on the whole, by both groups of listeners. This is in line with the difficulties of trial-by-trial talker variability which has been found to be detrimental in phonetic training of L2 pitch contours, particularly when the learner has weaker perceptual abilities (Perrachione et al., 2011). Wiener et al. (2018) investigated spoken word recognition in English learners of Mandarin using an eye-tracking paradigm, where learners were trained on single or multiple talkers. Eye-tracking data showed that learners in the low variability condition were less likely to look at the distractor items and generally required less time to switch from a more probably competitor to a less probable target. Furthermore, Antoniou et al. (2015) investigated the effect of talker variability in a word monitoring task with either single or multiple talker sentences, comparing native listeners to two groups of non-native listeners with different levels of previous language exposure. Non-native listeners with less exposure were slower and less accurate than native speakers regardless of the input

variability, but for the non-native listeners who had had more previous exposure, they were only weaker than the native listeners when the items were spoken by multiple talkers.

In summary, there is evidence that speech from multi-talkers is harder to process than single-talker speech, with evidence from spoken word processing, word recall, and for tone identification tasks. This is true for both native and non-native speakers, with the latter sometimes being more affected. This difficulty of multi-talker speech is in line with the finding that performance in the training stage has been found to be weaker in L2 phonetic and vocabulary training. While greater processing difficulty does not lead to weaker learning by necessity, there are claims in the literature that potential benefits of using multi-talker input in training (i.e. in terms of generalisation) might be offset by the higher processing cost of such input.

Rationale for the current replication study

From this literature review, we conclude that neither the seminal studies (Logan et al., 1991; Lively et al., 1993) nor many of the follow up studies provide conclusive evidence for a benefit of multiple talker over single talker input in phonetic training for generalisation abilities. There are only a small number of conceptual replications of this effect in the area of phonetic training as well as in other related areas, often reporting mixed findings. However, we have also seen that there is evidence that talker variability in training can be detrimental, with some studies arguing that there may be a trade-off in terms of the benefits of variability. This review therefore suggests that the claim that multiple talker input is generally more beneficial than single talker input for phonetic training, is not in fact well-established. On the other hand, the high variability benefit hypothesized by Logan and colleagues is intuitively sensible and we have seen that this early claim is in line with theoretical and computational accounts of language learning that have emerged over the last decades.

One difficulty in interpreting the mixed findings is that where null findings are reported, only Dong et al. (2019) and Brekelmans et al. (2020) have incorporated statistical methods (Bayes Factors) which allow us to evaluate the evidence for the null. Other studies with null

results used frequentist p-values as their inferential statistic, which means we are not able to conclude whether those studies that fail to find evidence for a high variability benefit actually provide evidence for the null – they may simply be underpowered to detect an effect. Notably, the majority of these studies did not provide power analyses, and samples are generally rather small, making it highly unlikely that the sample sizes provided are sufficiently powered to detect differences in learning across condition. An additional problem is the "file drawer problem", where null results may not be published as many journals have been reluctant to publish non-significant results in the past. This problem was already identified forty years ago (Rosenthal, 1979) but was still shown to be prevalent more recently (Ferguson & Heene, 2012; Masicampo & Lalande, 2012). It is thus possible that additional studies that have not found evidence for a benefit of multi-talker input have not been published. In the face of mixed evidence from a limited number of studies with small sample sizes, there is a strong motivation for 1) a high-powered replication in this area, and 2) for using an inferential statistic which allows for evaluation of evidence for the null.

For the current study, we therefore performed such a large-scale replication of the two key experiments in the original seminal studies of Logan et al. (1991) and Lively et al. (1993). We aimed to test the claim that learners generalise more after having had phonetic training on multiple talkers. We tested two groups of Japanese speakers who received either high variability (HV; stimuli spoken by five talkers) or low variability (LV; stimuli spoken by one talker) phonetic training on the /r/-/l/ contrast, following the design of the original studies. We compared improvement on a pre/post-test paradigm (a 2AFC identification test) with test items involving both trained and untrained talkers. Given the importance of being able to interpret any null finding, we used Bayes Factors as our key inference statistic to test our prediction, enabling us to quantify evidence for the null as well as for H1 (see Dienes, 2014). The use of this statistic also allowed us to use an efficient optional stopping procedure (Dienes, 2016; Rouder, 2014) with a planned minimum of 60 and maximum of 160 participants, as indicated could be required by our sample size calculations. Note that even

with the minimum value this is a substantial increase from the original sample size of 12 participants.

One possibility, given the literature reviewed above, is that we may *not* see the predicted overall benefit of variability, but that there is nevertheless a benefit for some subset of higher aptitude learners. Note that testing this claim is beyond the core goals of this work because our sample size calculations are based on establishing the two-way interaction between variability and pre- to post-test improvement, not a three-way interaction which would be substantially more difficult to power. Nevertheless, we include various measures of individual aptitude allowing for exploratory analyses which look at how these may modulate the benefits of phonetic training, which may be informative for future confirmatory work in this area.

Predictions

Our predictions – which we test in our confirmatory analyses – were as follows:

As our key test of the hypothesis that HV training aids generalisation of learning, we predict that:

1) Participants trained on HV input will show greater improvement on pre- to post- test items involving a novel talker and novel items than participants trained on LV input.

We test this by looking for evidence for an interaction between test-session (pre versus post) and variability-condition (HV versus LV). Note that, since this is our primary hypothesis, our sample size estimates are based on obtaining a sufficient sample for this effect.

We also test the following prediction, which reflects the observations made by Lively et al. (1993) concerning the difference in generalisation across their experiments:

2) Participants in the LV training set will show higher performance in a post-test with trained voices than in a post-test with untrained voices, while this difference will be absent or weaker in HV.

We test this by looking for evidence for an interaction between post-test type (post trained voice versus post untrained voice) and variability-condition (HV versus LV). Note that though we test this in light of the claims of the previous paper, we do *not* regard this prediction as the key test of the hypothesis that high variability training drives generalisation. The reason is that the interaction could be driven equally by weaker generalisation with the untrained voice in low variability participants *or* stronger performance with the trained voice by low variability participants (due to their greater opportunity to adapt to this speaker after having had more exposure to it in training).

We also test the following prediction about performance during training:

3) Participants trained with a single talker will improve more during training than those trained with multiple talkers due to stronger adaptation to the trained voice.

Although this is not reported in the original Logan et al., (1991) and Lively et al., (1993) experiments, as discussed above, a single talker benefit during training has been seen in other phonetic training studies (cf. Evans & Martín-Alvarez, 2016; Evans et al., 2017; Giannakopoulou et al., 2017), including in Dong et al., (2019) and Brekelmans et al., (2020) which used similar blocked approaches to variability in order to remove the difficulty of trial-by-trial adaptation. The hypothesis of a single talker benefit during training is tested to shed light on (2) and to help to interpret that result.

In addition, we test the following pair-wise comparisons in order to shed light on prediction 1 and 2:

4) Participants will show above chance pre to post improvement in both conditions, as has been seen in previous studies.
5) Participants in both conditions will perform better in the trained items post-test than in the untrained items post-test. (Note that although this prediction is unclear for HV, it is nevertheless set up as a hypothesis to be tested).

We also include measures of individual differences (e.g. attention, auditory discrimination ability). We do not have particular predictions related to these measures and investigate the patterns of results in exploratory analyses only.

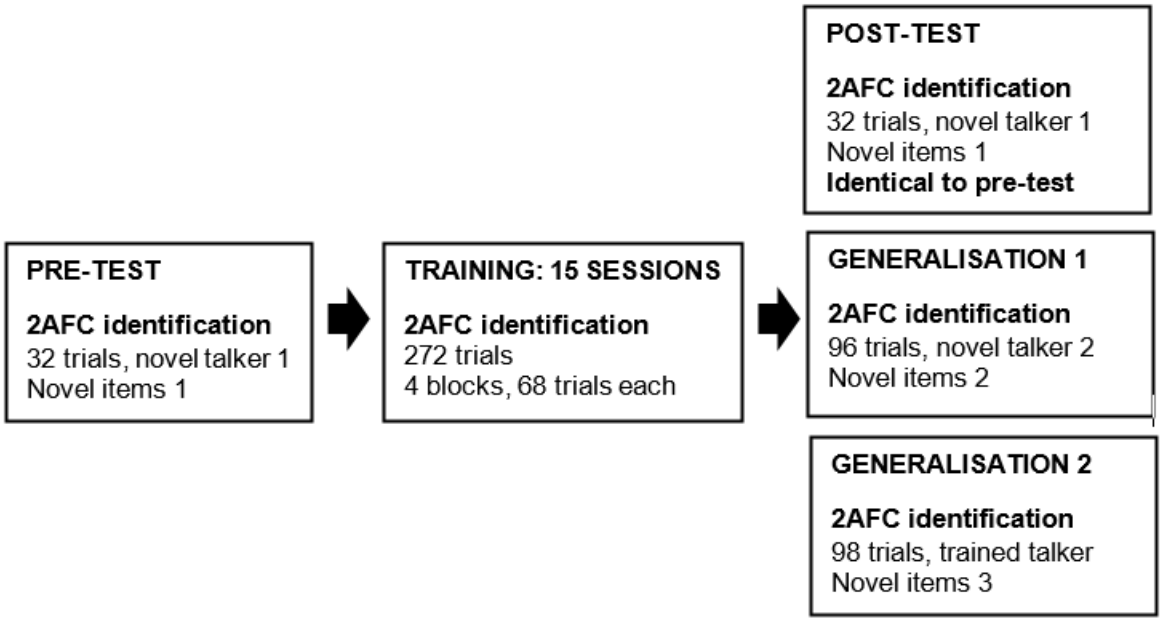
Methods

This study aims to replicate seminal research investigating the existence of a high variability effect. Our replication is intended to be as direct as possible given the details available in the original studies: Due to missing information, we did, however, have to infer some methodological details, such as participants' background details and language exposure, and information on the accent of talkers used to produce the stimuli. Further changes were made to address methodological issues in the original studies: One such crucial change affected how we draw conclusions from the pre/post-tests, as the original set-up does not allow for answering the key question regarding the effect of variability on generalisation. Details regarding this change are given in the design section below, and an overview of the experimental design of the original studies and this replication can be found in Figure 1. The procedure of the pre/post-test tasks itself is otherwise as direct a replication of the original design as possible (taking into account advances in computerisation). Another difference between our replication and the original studies can be found in the participant sample: While the original studies recruited Japanese participants who were living in the United States and were enrolled at a university, we instead recruited participants who are native Japanese speakers who are learners of English, regardless of their current location, in order to facilitate recruitment of the larger sample we require. Note that the original study used opportunity sampling and did therefore not select their participants based on their prior exposure to English, with no detailed data on this exposure being reported. This lack of information made it difficult to know what level of English exposure we should target in our participants. To ensure a coherent sample that is controlled for English exposure, the current study instead selected participants based on their pre-test score (details below). Additionally, we recorded new stimuli, as the original recordings of the stimuli are no longer available, according to the

original authors. To mitigate any talker-specific effects in this new sample of talkers (see e.g. Xie et al., 2021), we also counterbalanced the talkers across the tasks – which was not something that was done in the original. Finally, our statistical analyses differed substantially from the original (details below). This is due to statistical advances resulting in methods that are better suited for answering the key questions that were posed in the original studies.

In sum, our study is a direct-as-possible replication with changes only implemented where necessary or to better answer the original questions of interest. It should be noted that failing to find evidence in support of the key hypothesis could be due to some of the above changes, e.g. in population or talker recordings. However, if we should find that the original effect hinged on these factors, this would call into question the generalisability of the original claims, and thus would still be an informative outcome.

A: Original studies



B: Current replication

Figure 1. Overview of the design of the Lively et al., (1991) and Logan et al., (1993) studies (panel A), and the current replication (panel B). The numbers after certain talkers indicate distinct novel talkers (i.e. novel talker 1 is different from novel talker 2, but both are novel compared to the talkers used in training). Similarly, the numbers for the novel item sets

indicate distinct sets of items (i.e. novel items 1 is different from novel items 2 and 3, but all are novel compared to the items used in training).

Sample size estimation and effect size derivation

Sample size estimation

With original sample sizes being only 6 participants in each of the seminal experiments, we increased this sample size substantially in our replication. As we make use of Bayes Factors as our core method of inference, it is possible to test participants until we reach a pre-specified threshold for a given effect, since Bayes Factors remain a valid measure of the evidence even with optional stopping (Dienes, 2016; Schönbrodt & Wagenmakers, 2018)[1]. Further details on our use of Bayes Factors as inference criteria can be found in the Analysis section below.

With respect to our core hypothesis – predicting greater pre-post improvement in the high variability condition than the low variability condition – we set the following stopping criteria: We would periodically calculate the Bayes Factor for this hypothesis once participants had completed both pre- and post-test, starting from a minimum sample of N=60 (30 per training condition) up to a maximum sample of N=160 (80 per condition). Importantly, we would stop recruitment if there was *either* strong evidence for H1 (BF >10) or substantial/moderate evidence for H0 (BF <1/3). Any datasets from participants who had already started (but not yet completed) the experiment at the point at which we stopped recruitment, would be included in the final analyses. Note that our asymmetric criteria for null versus H1 reflect the (far) greater difficulty of obtaining evidence for the null as exemplified in the results of the computational simulation in Figure 2 below.

[1] We acknowledge that there has recently been some debate concerning whether optional stopping with Bayesian inference methods is problematic (de Heide & Grünwald, 2020). We follow recommendations from Rouder (2014) and Rouder & de Haal (2019), who point out that this is only the case when unprincipled values (i.e. defaults) are used to inform H1, which is not the case in this study.

Determining the minimal planned sample (N = 60): In principle, for the Bayes Factors, a minimal sample is not essential. However, having a reasonably-sized minimum sample is essential to maintain the validity of coefficient estimates and standard error sizes of our models: These are obtained through the logistic mixed model regression, where very small samples can lead to biases in these estimates. In light of this, we set a sample of 30 participants per condition as our minimum sample. The rationale behind this minimal sample size is that this sample is still larger than the largest sample size per variability condition in any phonetic training study with adult participants (as far as we are aware), which is 29 as used in Shinohara (2021).

Determining the maximal planned sample: Our goal here was to determine a maximal sample size for our key hypothesis which would be large enough to find evidence for H1 or H0 respectively, depending on which was actually true. This approach thus minimizes the chance of finishing data collection with ambiguous data (i.e. finding no evidence for either H1 or H0). To find this sample, we simulated 1000 random data sets with 60 binary responses at pre- and 60 at post-test for each participant (as in our experiment) for different values of *N* (total number of participants). We did this for two scenarios: where H1 is true, and where H0 is true.

For the simulation where H1 is true, we needed an estimation of the relevant parameter for the size of the interaction between variability and test session, i.e. our predicted effect size. We set this to odds ratio 1.37 (0.32 in log odds). Although the original studies do not report this effect size directly, we computed it based on observed performance in the experiments reported across Logan et al. (1991) and Lively et al. (1993).

The logic for the calculation of the odds ratio is as follows:

Estimation of average odds of picking correct option (out of two options) at pre-test: We estimated this from the average performance at pre-test across the two experiments reported by Logan et al. (1991) (on the basis that any differences between the two are accidental): this is 73.42% or odds = 2.76.

Estimation of odds of picking correct option (out of two options) at post-test after HV training: We estimated this from the average performance at post-test reported in Logan et al. (1991): this is 82.7% or odds = 4.78.

Estimation of odds of picking correct option (out of two options) at post-test after LV training: We estimated this from the average performance in the two post-tests with untrained voices in Lively et al. (1993): this is 77.68% or odds = 3.48.

Estimation of odds ratio for change from pre- to post-test following HV training: 4.78/2.76 = 1.73
Estimation of odds ratio for change from pre- to post-test following LV training: 3.48/2.76 = 1.26
Estimation of odds ratio for HV over LV benefit for change from pre to post: 1.73/1.26 = 1.37

Note that this same value is also used when computing the Bayes Factor to test this interaction since, as explained in Analysis, Bayes Factors also require an estimation of the predicted effect size - Table 3 in the Analysis section lays out the predicted effect size for each hypothesis and shows how they were computed detail.

For the simulation where H0 was true, we set the relevant parameter to be odds ratio 1 (0 in log odds)[2]. For both scenarios, we ran a logistic mixed effects model, from which we

---

[2] Other parameters were fixed across both simulations as follows: overall difference between test sessions odds ratio = 1.47, overall difference between variability conditions = 1.15. These effect sizes were computed on the basis of mean percentage accuracies reported in Logan et al. (1991) and Lively et al. (1993) as follows: Average for high-variability group at post-test computed from the average performance in the two post-tests ('post-test' and 'TG1') with untrained voices and items in Logan et al. (1991) (82.7%). Average for low-variability group at post-test computed from average performance in the two post-tests ('post-test' and 'TG1') with untrained voices and items in Logan et al. (1991) (77.68%); average for both groups at pre-test computed from average performance at pre-test across the two experiments, i.e. assuming that any differences at pre-test were accidental in that sample (73.42%). No factors related to items or any other features of stimuli were accounted for in the simulations. For the estimation of random effects there is no relevant information provided in the original papers, nor in any available published paper with sufficiently similar methods to be appropriate. However, we were able to estimate these from a dataset kindly

extracted the estimate and SE for the interaction, and from these computed the Bayes factor and the predicted effect size (according to the procedure outlined under Analysis below). From the 1000 simulations, we then computed the proportion of the time that various BF thresholds are met (see Figure 2). Details of the simulations with analyses scripts and simulated datasets are available on the OSF: https://osf.io/4hkqb/

From the simulations it is clear that, even with small samples, it is very rare to find evidence for H0 when H1 is true, as indicated by the lines where BF<1/3, BF<1/6 and BF<1/10 hover around 0 in the left plot. It is similarly rare to find evidence for H1 when H0 is true, indicated by the lines where BF>3, BF>6 and BF>10 hover around 0 in the right plot. However, the evidence is often ambiguous (see the line where 1/3<BF>3). This is particularly apparent in the case where H0 is actually true. Ideally, we planned to choose as our maximum a value of *N* where we can obtain evidence for H1/H0 in 80% of samples. This is, however, not feasible for the case where H0 is true: even with 210 participants, we do not see 80% of samples meet even the most moderate criteria of BF<1/3.

Therefore, as a balance with feasibility, we decided on a final sample of 160 participants. If H1 is true, this gives 99% chance of chance of obtaining appropriate moderate evidence (BF>3) and 94% chance of obtaining strong evidence (BF>10)[3]. If H0 is true, it gives us 68% chance of obtaining moderate evidence for this, albeit only 7% chance of obtaining strong evidence (BF<0.1). The greater difficulty of obtaining evidence of H0 than H1 is also why we

---

made available to us by Anastasia Giannakopoulou (for the experiment reported in Giannakopoulou et al., 2014), and they were: Participant intercept: SD = 0.16; by-participant slope for test-session = .19; *r* = -.402).

[3] For comparison purposes, given their greater familiarity to readers, we also looked at p-values, specifically for the runs where H1 was true at N=160: 99% of samples, met criteria p<.05 while 92% met criteria p<.005.

had asymmetric stopping criteria: we would not stop collecting data even if we got substantial evidence for H1 because we know it could be feasible to collect a sample with strong evidence for H1, however we would stop collecting data when we have substantial evidence for H0 because we know it is unlikely we could obtain strong evidence in this direction.

Note that the interpretation of the Bayes Factor does not depend on power calculations: That is, our simulations here are for determining the feasibility in terms of resources for obtaining different strengths of evidence, and they do not influence how our final results should be interpreted.

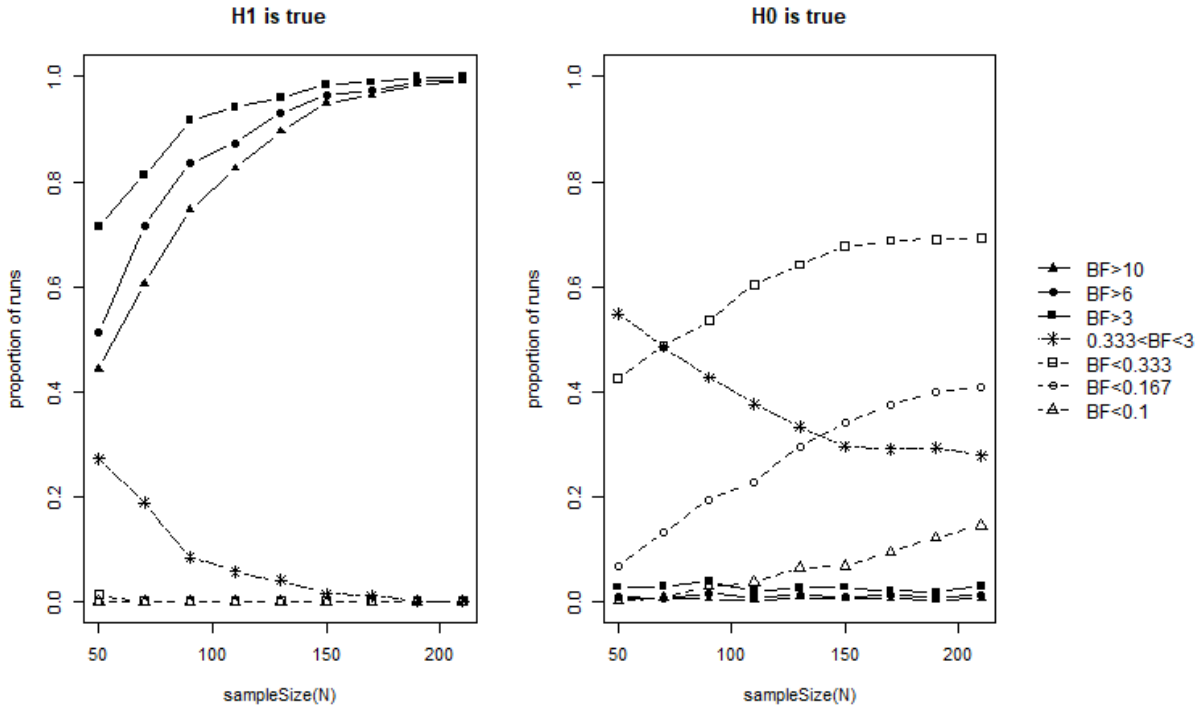


Figure 2: Results of simulations where H1 is true and H0 is true with different sample sizes (N) showing proportion of runs meeting different BF criteria for 1000 runs per sample size.

Participants

Participants were native speakers of Japanese who are learning English and are aged between 18-40 years. While the original studies recruited participants who were living in the United States at the time of testing, our participants were native Japanese speakers, but we

did not restrict recruitment based on their current location. In line with the original studies, participants were asked to rate their ability to understand written and spoken English. We further collected information on the amount of time participants have been learning English, whether they have spent any time living in an English-speaking country, whether they are currently enrolled in English language classes, and what percentage of time they communicate in English in daily life (full questionnaire available on the OSF). Participants were recruited through the online recruitment platform Prolific.co as well as personal contacts. They were paid for their participation in the study at a rate of at least £5 per hour, with a bonus for completing the full experiment and for timely completion. The study has been approved by the Queen Mary Ethics of Research Committee, Project ID Number QMERC20.365.

Exclusion Criteria: As in the original study, participants were excluded if they report any history of speech or hearing impairments, language disorders, or dyslexia. Additionally, anyone who moved to an English-speaking country before the age of 18 was excluded. Participants who failed a headphones screening (Milne et al., 2020) or those who failed catch trials aimed at checking attention and effort will also be excluded (as detailed in the Procedure below). Furthermore, those participants who were above our ceiling upper cut-off performance of 85% at pre-test were excluded: This upper cut-off point is based on the general pattern of improvement in phonetic training paradigms, which find around 10-15% improvement (see also Iverson et al., 2005 who used a pre-test score as a cut-off to avoid a ceiling effect). Finally, those who had a gap between sessions of more than three consecutive days were excluded. Excluded participants were replaced.

Sample Size: In line with the sample size estimation described above, our minimal planned sample was 60 participants (30 in each variability condition), and our maximal planned sample was 160 (80 in each variability condition).

Randomization: Participants were assigned to one of the two variability training conditions (HV and LV) to ensure as close to equal numbers in both conditions as possible. We aimed

to match the variability condition groups based on the participants' initial aptitude at perceiving the non-native phonetic contrast at pre-test (a 2AFC identification task, capturing their ability to identify /l/ and /r/-sounds prior to training – further details under Stimuli below). To do this, we counterbalanced the assignment of training conditions to participants based on their pre-test scores. To achieve this, we created two groups of participants ('high' and 'low' pre-test scorers) using an algorithm embedded in our experimental program. The initial dividing point was 70% as a middle ground between chance (50%) and our upper cut-off (85%), but this value could be adjusted if participant drop-out rates resulted in unbalanced groups. Within those two groups, we randomly assigned participants to the HV or LV condition.

Through this process, we therefore ensured that groups were as closely matched between the HV and LV condition as possible, and that groups were not at ceiling before starting the training (see Exclusion Criteria). Floor effects would have likely been less of a concern given our learners are adults with some English experience. Furthermore, a similar kind of phonetic training paradigm but with 7- and 11-year-olds who had very little English experience showed improvement over a set of 8 training sessions, even if pre-tests scores were at chance (Brekelmans et al., 2020). Given that 1) our learners generally had more experience of English and 2) there is no reason to suspect that phonetic training per se only works for certain pre-test accuracy, we did not include a lower cut-off for pre-test scores. Our recruitment criteria around participants' initial aptitude therefore differ from the original studies. We note, however, that while matching the recruited participant group precisely to the original studies may seem desirable, the difference of pre-test scores across conditions in the original studies is large (from an average of 69% in the 1993 LV paper to an average of 78% in the 1991 HV paper), information about the range of aptitudes is missing, and is only based on a small sample of 6 participants in each of these.

Design

The study used a pre/post-test design building on the design used in the original studies, and comprised three stages: pre-test, training, post-test. Each of the stages will be described in more detail below. As in the original studies, the pre-test task consisted of a two-alternative forced choice (2AFC) identification task. Training itself was also a 2AFC task, with the addition of trial-by-trial feedback. At post-test, participants did two 2AFC tasks similar to pre-test, but with different voices and items to be able to measure generalisation. Further additional tests of individual difference measures were added before the pre-test.

Note that a key difference between the original studies and our replication is in how we draw our conclusions about generalisation. In the original studies, an identical pre-test and post-test task were used which contained items and talkers not used in training as well as a set of filler items testing different phonetic contrasts beyond the /l/-/r/ contrast (e.g. deep-keep). This identical pre/post-test task is the task on which pre/post-test improvement was measured, however it is not the task that was used to test the crucial prediction about generalisation. The key tasks used to test differences in generalisation between the groups were two additional generalisation tasks after training: one that used untrained items and a new untrained talker, and a second that used different untrained items but a talker used in training. Crucially, the original studies' comparison of these generalisation tasks did not take the pre-test performance into account, such that differences in the generalisation performance across conditions could be due to the training or, in fact, to pre-existing differences in performance between participant groups at pre-test. To account for potential differences at pre-test (but see Randomization), we used the same pre-test and generalisation tasks as the original studies did, except that we did not include fillers, and, critically, we measured changes in performance between the pre-test and the two generalisation tasks. This means that improvement from pre- to post-test will truly have been due to training rather than possible pre-test differences, and differences in post-training performance between generalisation task 1 and 2 will have been due to generalisation

differences caused by talker familiarity. Because we are most interested in generalisation, we therefore dropped the post-test task that is identical to pre-test in both talker and items.

Talker variability was manipulated between participants, with half assigned to HV and half to LV training. For HV training, participants heard five different talkers while in LV training stimuli were spoken by a single talker (one of the five used in HV, rotated across versions). A sixth and seventh talker (i.e., ones not in training, again rotated across versions) were used for the pre-test and the first post-test task (Generalisation 1), while one of the familiar talkers (i.e., one of five used in training for HV, or the only one used in training for LV) was used in the second post-test task (Generalisation 2). See Table 1 for an overview.

VERSION	TALKERS HV TRAINING	TALKERS LV TRAINING	TALKERS PRE-TEST	TALKERS GENERAL-ISATION 1	TALKERS GENERAL-ISATION 2	ITEMS TRAINING	ITEMS PRE-TEST	ITEMS GENERAL-ISATION 1	ITEMS GENERAL-ISATION 2
1	trained 1/2/3/4/5	trained1	new1	new2	trained1	Training items	set3	set1	set2
2	trained 2/3/4/5/1	trained2	new2	new1	trained2		set1	set2	set3
3	trained 3/4/5/1/2	trained3	new1	new2	trained3		set3	set1	set2
4	trained 4/5/1/2/3	trained4	new2	new1	trained4		set1	set2	set3
5	trained 5/1/2/3/4	trained5	new1	new2	trained5		set3	set1	set2
6	trained 1/2/3/4/5	trained1	new2	new1	trained1		set1	set2	set3
7	trained 2/3/4/5/1	trained2	new1	new2	trained2		set3	set1	set2
8	trained 3/4/5/1/2	trained3	new2	new1	trained3		set1	set2	set3
9	trained 4/5/1/2/3	trained4	new1	new2	trained4		set3	set1	set2
10	trained 5/1/2/3/4	trained5	new2	new1	trained5		set1	set2	set3

Table 1. Illustration of the talkers and items used in each task across variability conditions. Talkers and items were rotated for counterbalancing, to ensure any effects particular to a speaker's voice were avoided. Talker numbers for talkers used in HV training indicate the order in which talkers were presented.

As in the original studies, talkers were drawn from a pool of 7 native speakers of North American English. While the original studies did not use counterbalancing but instead used a fixed talker for single-talker tasks, we counterbalanced the assignment of talkers and items to avoid unwanted talker- or item-specific effects (see e.g. Xie et al., 2021). In the HV condition, stimuli from the five talkers were presented on separate days in line with the original studies. Talker order was counterbalanced across days. Further counterbalancing can be found in Table 1. As fully counterbalancing both talkers and items across all tasks would lead to 90 counterbalanced versions per variability condition, we instead counterbalanced the talkers within tasks by ensuring the two new talkers in pre/post-test never matched, and ensuring that all trained talkers were used as the trained talker in Generalisation 2. Sets of items were varied across the tasks, such that one half of the versions used list 1 for pre-test, list 2 for Generalisation 1 and list 3 for Generalisation 2, and the other half used list 3 for pre-test, list 1 for Generalisation 1 and list 2 for Generalisation 2. This lead to 10 counterbalanced versions per variability condition, to ensure that any differences in improvement could not be due to characteristics of the individual talkers or items (i.e. one talker being significantly more intelligible than others, or certain items being easier/harder than others).

Stimuli

Stimuli were recordings of spoken minimal pairs contrasting /r/ and /l/ in word-initial and word-final positions, in singleton and cluster environments, and in intervocalic positions (e.g. ***r****ock* for initial singleton, *f****l****ock* for initial cluster, *pi****l****ot* for intervocalic, *wa****ll*** for final singleton, and *ha****r****d* for final cluster). Talkers were seven monolingual native speakers of Canadian English (5 female, 2 male). Stimuli were recorded in sound treated booth using a head mounted microphone at a sampling rate of 48,000 samples per second. They were high-pass filtered at 65 Hz to eliminate potential background noise, and then downsampled to 24,000 Hz. Finally, intensity was scaled to 70 dB. All processing was carried out in Praat (Boersma & Weenink, 2015).

The stimuli used across the tasks were primarily items that were used in the original studies. Some items form the original studies were replaced with items fitting the same criteria: we replaced items in cases where the original items were parsed as non-words by a native speaker, or where the original item pairs did not result in minimal pairs that only differed in the /l/-/r/ contrast in the accent variety of the speakers we used to record the stimuli (see Appendix C for more details on this for specific stimuli items). In line with Logan et al. (1991), stimuli validity of the specific tokens was established by testing each token for intelligibility in an open-ended identification task on a group of 160 native English speakers, where listeners were asked to type the word they heard. In this intelligibility task, each participant was presented with 161 individual tokens spoken by a number of the recorded talkers, and a total of 16 different stimuli sets was created. In this way, the intelligibility of each token for an individual talker was assessed by 10 different native English participants. Tokens that were not intelligible to at least 80% of the participants in terms of having errors in the perception or /l/ or /r/ were replaced by different tokens from the same talkers. One talker was replaced as seven of their tokens had an /l/-/r/ accuracy below 80%. One token which had an /l/-/r/ accuracy of 78% was replaced among the other talkers.

Stimuli lists have been made available in Appendix C in written form, and all stimulus recordings are available on the Open Science Framework. Since word frequency varies between the two items in a minimal pair, word frequency of the individual stimuli items was gathered from the Corpus of Contemporary American English (Davies, 2008). Similarly, we collected familiarity scores for the words in the minimal pairs as part of the individual difference measures. These data may be used in an exploratory analysis to examine the effects of frequency and prior knowledge on phonetic training outcomes.

Training stimuli

As in the original studies, 136 individual stimuli items were used in training, presented as 68 minimal pairs. These minimal pairs contrast /r/ and /l/ in five phonetic contexts: word-initial pre-vocalic singleton (***r****ock*), word-initial clusters of a consonant and the liquid (*fl**l**ock)*, word-

medial intervocalic (*pi**l**ot*), word-final post-vocalic singleton (*wa**ll***), and word-final post-vocalic cluster (*ha**r**d*). The spread of items across contexts was identical to the original studies. Each minimal pair was presented twice, to ensure both items in the pair occur as the target and the competitor throughout each training session. In addition, stimuli were repeated once, in line with the original studies, totalling 272 trials in each training session.

Pre/post-tests

There was one pre-test, which used a novel talker and a novel set of items not used in training. After training, there were two post-tests, that test for different aspects of generalisations: Generalisation 1 used a different novel set of items and a different novel talker, while Generalisation 2 used yet another novel set of items but a talker that was also used in training. Thus, none of the items used in pre-test or either of the post-tests were used in training, ensuring any effects of variability were not item-specific.

There are some differences between the current study and the original study in relation to these pre- and post-tests. In the original studies, the pre/post-test used 32 stimuli items presented as 16 pairs, with an additional 8 filler pairs that differed in other contrasts and were not analysed. The original Generalisation 1 used 96 stimuli items presented individually, i.e. not all items had a viable minimal pair contrasting the /r/-/l/ contrast, and the original Generalisation 2 used 99 stimuli items once more singly presented with not all items having a viable minimal pair item. In our study, we have streamlined the pre/post-tests to ensure equal numbers of stimuli items in each task, and to ensure all stimuli items were plausible English words with both /r/ and /l/ used in the critical phonetic context. Each of these tasks used 60 individual stimuli items presented as 30 minimal pairs. These minimal pairs again contrasted /r/ and /l/ in the same five phonetic contexts as used in training. Each minimal pair was again presented with both items in the pair occurring as the target and the competitor.

Procedure

The study was run in Gorilla (www.gorilla.sc, Anwyl-Irvine et al., 2019). Participants were asked to complete the experiment in a quiet environment and completed a screening to ensure they were wearing headphones (Milne et al., 2020). As part of the headphone screening, participants were able to adjust the stimulus volume to a comfortable listening level. Throughout all training sessions, we incorporated catch trials as attention checks. These trials played a sentence asking participants to click a specific button ('Please click the button on the left/right'). If participants failed more than 20% of the attention checks in the first training session (i.e. more than 3 of the 16 catch trials), participants were excluded. Any analyses regarding differences in attention (as indicated by the number of correct responses to the catch trials) throughout the rest of the training sessions would be considered exploratory.

An overview of the procedure can be seen in the flowchart of Figure 1. Participants first completed the measures of individual differences and the pre-test between five days and 24 hours before their first training session, and then undertook 15 training sessions, as in the original studies, each on a different day. The post-test, consisting of two generalisation tasks, was completed between 24 hours and five days after the last training session, ensuring that any temporary phoneme categories built up during training were no longer maintained (Heeren & Schouten, 2010). The generalisation tasks at post-test were identical in procedure to the pre-test task. The entire experiment took around a month at most to complete, with no more than three consecutive days between sessions. Both accuracy and response times were collected for every trial, with response latency measured from the onset of the stimulus presentation.

Training

The training task is identical to the original studies and is a 2AFC identification task. In each trial, participants heard a word that is part of a minimal pair that contrasts /r/ and /l/. Participants saw two buttons on the screen indicating the target word and the competitor

(i.e., the minimal pair of the target). Their task was to select the word they think they had heard. Participants then received feedback as to the accuracy of their answer, and were played the word again and shown the correct response. Training consisted of 15 sessions, comprising 68 minimal pairs presented twice with each of the 136 training words as target, as in the original studies. In total, 272 trials were presented per session in four random blocks of 68 trials. Trial order was randomised across participants. During each training session, stimuli from only one talker were presented, as was the case in the original studies. Participants in the HV condition cycled through the set of five training talkers three times throughout the 15 training sessions. Participants in the LV condition completed the training session with a single talker 15 times. Each session lasted approximately 20 minutes each, and participants were able to track their progress using a progress bar.

Pre/post-test tasks

The pre-test and the two post-test tasks were identical to the training task, except no feedback was provided. As such, participants were presented with 30 minimal pairs (60 individual trials) contrasting /r/ and /l/. In these 2AFC identification tasks, participants chose the word they had heard via response buttons showing the written minimal pair items presented on screen, as in the original studies. The pre-test task lasted approximately 10 minutes. The two generalisation tasks at post-test tasks taken together lasted approximately 20 minutes.

Individual differences

To further explore potential effects of other factors on any patterns of results we might find, several tests of individual differences were added at pre-test. These included tests of auditory processing ability, attention, and familiarity with the words used in the study (see Table 2 for an overview), as well as a background questionnaire (available on the OSF). The combined individual differences took approximately an hour to complete. Any analyses involving these tests would be exploratory.

Measure	Tasks	Source/Reference
Auditory processing	- Duration discrimination - Formant discrimination - Frequency discrimination - Melody memory - Rhythm memory - Rise time discrimination	Measures described in: Saito, K., Suzukida, Y., Tran, M. and Tierney, A. (2021). Reliability of the measures reported in more detail in this preprint: Saito, K., Sun, H., and Tierney, A. (2020, 12 June).
English vocabulary	Lextale	Lemhöfer, K., & Broersma, M. (2012).
Familiarity	Familiarity rating task	Purpose-built
Attention	1. AXCPT 2. Dichotic listening 3. Selective auditory attention	Adapted from: 1. Cohen, J.D., Barch, D.M., Carter, C.S., & Servan-Schreiber, D. (1999). 2. Koch, I., Lawo, F., Fels, J. & Vorlaender, M. (2011). 3. Woods, K. J., & McDermott, J. H. (2015).

Table 2. Detailed overview of the individual difference measures employed at pre-test.

Analysis

Data exclusion and pre-processing

Participants who met any of the exclusion criteria specified in the Participants section above were excluded from data analysis. Those who failed the headphones screening after a chance to retake it, or those who failed the catch trials, were also excluded, as were those who did not complete the entire set of training sessions or failed to return for the post-test.

Specific trials that had a response time of more than 3 standard deviations from the participant mean were also excluded from analyses. Participants' responses were coded as trialwise accuracy (1 - correct, 0 - incorrect).

Inference criteria - Bayes Factor analyses

We tested a series of targeted hypotheses - as laid out under *Predictions* at the end of the Introduction and in Table 3 below - using the Bayes Factor (BF), which computes the strength of evidence for the hypothesis (H1) over the null hypothesis (H0), or vice versa. This provides a measure of the strength of evidence for a hypothesis (H1), compared with the null (H0). These differ from p-values which do not provide the option to evaluate the strength for the H0 ($p > .05$ does not tell us that we should increase our confidence in the null hypothesis, despite this common misinterpretation). Bayes Factors also have the advantage that they provide a continuous measure of the level of evidence. Since Bayes Factors might be less familiar, *p*-values are additionally reported for the reader's convenience. We will, however, only interpret the results with respect to the Bayes Factors.

Bayes Factors were computed using the method advocated by Dienes (2008, 2014, 2015), modelling H1 as a half normal, testing a one-sided prediction in each case (appropriate since only directional hypotheses are tested). Note that using a half normal rather than a uniform distribution has the advantage of favouring smaller values (Dienes, 2014). This method of computing Bayes Factors requires three numbers: (i) an estimate of mean difference, (ii) the standard error (SE) (iii) the prior – an estimate of *x*, the predicted mean difference under H1, which is then used as the standard deviation of the half normal modelling H1. (i) and (ii) comprise our data summary. We obtained these by running generalised logistic mixed effects models (glmer; Baayen et al., 2008) using the lme4 package (Bates et al., 2015) for the R environment (R Core Team, 2018) and extracting beta and SE values from the relevant coefficients. Note that these values were in log odds space, and this approach allows us to meet the assumptions of normality required for this approach to computing the Bayes Factor. Our approach to building the relevant glmer models is described more fully below.

A key issue in Bayes Factor analyses is how to determine an appropriate value for (iii), i.e. our estimation of the predicted effect size *x* (also called the prior). We estimated these based on previous data as outlined in Table 3. In most cases, the data are computed using values from the relevant experiments in Logan et al. (1991) and Lively et al. (1993). Where we test for differences within training, since these are not reported in the original studies, we use values from Dong et al. (2019). Our choice of values of H1 for each test, and the justification, is laid out in Table 3 below. Bayes Factor results are reported using the notation $BF_{H(0,x)}$ to denote a Bayes Factor where *x* is the standard deviation of the half normal used to model H1 (following Dienes, 2021).

Bayes Factors were interpreted as a continuous measure of the evidence but also with reference to the conventions in Jeffreys (1961) and expanded by Dienes (2014), where BF > 3 indicates moderate/substantial evidence for H1 and a BF >10 indicates strong evidence for H1, a BF between 1/3 and 3 indicates ambiguous evidence for H1 (the data is insensitive to test the hypothesis), a BF < 1/3 indicates moderate/substantial evidence for H0 and a BF < 1/10 indicates strong evidence for H0[4]. Since there is subjectivity in the choice of value to inform H1, robustness regions are included, showing the range of estimates of H1 (i.e., the value used as the SD of the half normal) for which our data would support the same basic conclusion. Where we find evidence for H1 (either moderate or strong) we report the range of estimates which would also yield at least moderate evidence for H1 (i.e. a BF>3). Similarly, where we find evidence for H0 (either moderate or strong) we report the range of estimates which would also yield at least moderate evidence for H0 (i.e. a BF<1/3). This is noted as [x_1, x_2] with x_1 being the smallest SD and x_2 the largest SD (as recommended by Dienes, 2021). To compute the ranges, values were tested in increments of 0.01 from a difference of 0 from chance to the log-odds score corresponding to the difference between chance and ceiling

[4] Note that Bayes Factors provide a *continuous* measure of the evidence. Therefore – although we make use of the criteria in our stopping procedure – exact values will be reported and interpreted. This is a key difference in the interpretation of Bayes Factors and frequentist statistics (i.e. p-values).

performance (deemed to be all but one trial correct, i.e., 99.6% (log-odds 5.52) in Training; 97.9% (log-odds 3.84) in Pre/Post-task; 99.0% in both Generalisation tasks (log-odds 4.60). Note that there will always be some value which provides evidence for H0. Where this was not found within the range tested, the end point of the tested range is noted as e.g., >5.52 (Training), except for ranges of values giving evidence for H0, where the maximum is infinity. Robustness Regions should be interpreted bearing in mind that larger values of H1 bias evidence for the null, whereas smaller values bias in favour of H1.

Model building

As noted above, for each hypothesis test we extracted the size of the difference (beta) and SE from a logistic mixed effect model. Our approach to building these models was to include all experimentally manipulated variables and their interactions for each task as fixed factors in the models, regardless of whether they contributed to the model. Key fixed effects were *variability* (HV/LV), *session* (1-15 for training), *test* (pre or post for the pre/post-test), *talker novelty* (in generalisation only) and all the interactions between them.

The approach to random effects was to automatically include participant as a random effect with a full random slope structure – or the most complex to converge – as recommended by Barr et al. (2013). For item, random intercepts were automatically included, however, by-item slopes were only included if they contributed to the model following the model comparison approach recommended in Matuschek et al. (2017). The final model structure detailing the random effects structure and included control variables is reported for each model.

A simple contrast coding was used[5] for most fixed effects (*condition*, *test*, and *talker novelty*), where each level of a contrast is compared to the reference level with the intercept corresponding to the grand mean. *Session* was coded as a continuous numeric factor centred on 0. This coding was used to reduce collinearity effects between main effects and

[5] For the detailed code, see the functions *myCenter, lizCenter,* and *lizCenter2* on the Language Learning Lab Github page at https://n400peanuts.github.io/languagelearninglab/usefulFunctions-explained.html

interactions, and so that the intercept corresponded to the grand-mean (in log-odds space) and main effects were evaluated as average effects over all levels of other factors. All models that are reported converged using the Bound Optimization by Quadratic Approximation (BOBYQA optimization: Powell, 2009), unless specified otherwise.

In addition to the experimental factors, factors to control for phonetic context of the contrast (5 contexts: word-initial pre-vocalic, stop + liquid cluster, word-medial intervocalic, and word-final postvocalic), and talker (5 talkers in training, 3 in the tests) were included in the analyses but only if they contributed to the model fit (with $p > .2$ similar to what is suggested for random effects by Matuschek et al., 2017). These factors were not interpreted but were included (if meeting criteria) as it was hypothesized that they would likely contribute variance to the data (different phonetic contexts affect the acoustics of the consonants of interest, word familiarity and talker idiosyncrasies contribute to difficulty). Conceptually, talker and phonetic context are random effects, but these cannot have fewer than 6 levels (Bolker, 2020) and thus both were included as fixed effects in the model coded as a set of simple-coded contrasts (for phonetic context - 5 levels in both the pre/post-test tasks and in training; for talker – 4 for training and 2 for the tests) so that the intercept continued to represent the grand mean. Neither were included in any interactions.

Importantly, although full models were built (thus controlling error across the various factors) our key analyses only considered the targeted coefficients[6] relevant to our hypotheses, as laid out in Table 3 below. Inspection of any other coefficients would be considered exploratory. In addition, further analyses may also include the individual difference measures as predictors, and these would all be considered exploratory.

[6] We report the beta and SE values as well as the Bayes Factor computed using these values. We also report z and p values which are automatically computed by glmer. As noted above, care must be taken with these frequentist values given our optional stopping procedure, and we have not interpreted them but include them due to their familiarity to the reader.

Contrasts to be tested

Table 3. Predictions that were tested, with the motivation for testing this prediction, with the estimated effect size for each prediction. For each prediction, the logistic mixed effects model from which the summary data was used is indicated.

			SUMMARY DATA		MODEL OF H1
	PREDICTION	MOTIVATION	MODEL FROM WHICH COEFFICIENT STATISTICS ARE EXTRACTED	RELEVANT COEFFICIENT (from which beta and SE are extracted)	ESTIMATE OF PREDICTED EFFECT SIZE *x* IN LOG ODDS. This will be set as SD of half normal with mean of 0. Odds ratio in parentheses.
1	Participants in the HV condition improve more from pre-test to post-test 1 (untrained talker	Exposure to multiple talkers (HV) boosts generalisation in phonetic training compared with exposure to a single talker (LV)	*model 1:* predicting accuracy in	test-session by variability condition	0.32 (1.37)[1]

	and untrained items) than do participants in the LV condition	[Note: we regard this as the key test of HV benefit]	pre-test data + post-test1	LV and pre-test at reference level*	
2	Participants in the LV condition show higher performance in post-test 2 (trained talker) than in post-test 1 (untrained talker) but this difference is absent or weaker in HV	Exposure to a single talker (LV) in training will lead to adaption to that voice, giving an advantage when tested with that voice. In contrast, multi-talker (HV) training will lead to learning in a more general, less talker-specific way [Note: this is included as the test most closely corresponding to the claims made in Lively et al., 1993]	*model2:* predicting accuracy in post-test1 data + post-test2 data	test-session by variability-condition HV and novel talker at reference level	0.27 (1.31)[2]

3a	Participants in LV will show a steeper learning slope during training than participants in HV	Repeated exposure to a single talker (LV) in training will lead to gradual adaption to that talker over sessions, which cannot occur to the same extent when training is divided over multiple talkers [Note: this is not reported in Logan et al, but has been found in other studies in the literature (e.g. Brekelmans et al., 2020; Giannakopoulou et al. 2017; Dong et al., 2019). If we see a benefit in LV, this may help us to interpret findings for prediction 2]	*model3a:* predicting accuracy in training data	variability-condition and variability-condition by training-session HV at reference level	0.35 (1.42) [3a] and 0.8 (2.25) [3b]

3b	In training, from session 2 onwards, participants in LV will perform better than participants in HV	From the second session, participants in LV are exposed to a talker they have previously heard in each session, whereas this is not the case in HV [Note: this is closely related to 3a, and is included for the same reasons, and <u>in addition</u> because we have previously found power to detect the main effect of variability where we cannot detect the interaction with session]	*model3b:* predicting accuracy in training data (with session 1 removed)	variability-condition HV at reference level	0.35 (1.42) [3a]

4	Participants in (1) HV condition and (2) LV condition will both improve from pre-test to post-test 1 (untrained)	Pre- to post-test improvement was seen both *in Logan et al*. (1991) and Lively *et al*. (1993), and in many other papers [Note: these tests are not key to our hypothesis but are included since they will generally help us understand performance in the experiment and shed light on factors underpinning prediction 1]	*model 4:* versions of model1 with HV and LV coded as the reference level	test-session (we will be looking at this effect at each level of variability-condition) Pre-test at reference level	0.45 (1.57)[4]
5	Participants in (1) HV condition and (2) LV condition will perform better in post-test 2 (trained)	Mean for post-test with trained talker was numerically higher than post-test with untrained talker in both HV and LV in Logan et al. (1991) and Lively et al. (1993) *[Note: these tests are not key to*	*model5:* versions of model2 with HV and LV	test-session (we will be looking at this effect at each level of	1.17 (0.16) [5]

	than post-test 1 (untrained)	*our hypothesis but are included since they will generally help us understand performance in the experiment and shed light on factors underpinning prediction 1]*	coded as the reference level	variability-condition) Novel talker at reference level	

* Note the fixed effect of *variability* will be given a reversed coding in model 1 , so that in each case a positive beta for the interaction will be in the direction of the prediction

[1] Computed from means from Logan, et al. (1991) and Lively, et al. (1993) as follows:

odds of choosing correct option at pre-test: 2.76

odds of choosing correct option for HV in untrained talker post-tests: 4.78

odds of choosing correct option for LV in untrained talker post-tests: 3.48

odds ratio for HV for change from pre to post-tests: 4.78/2.76 = 1.73

odds ratio for LV for change from pre to post-tests: 3.48/2.76 = 1.26

odds ratio for HV over LV benefit for change from pre to post : 1.73/1.26 = **1.37**

[2] Computed from means from Logan, et al. (1991) and Lively, et al. (1993) as follows:

odds of choosing correct option for LV in trained talker post-tests: 4.88

odds of choosing correct option for HV in trained talker post-tests: 5.13

odds of choosing correct option for LV in untrained talker post-tests: 3.48

odds of choosing correct option for HV in untrained talker post-tests: 4.78

odds ratio for HV for benefit of trained over untrained talkers at post-test: 5.13/4.78= 1.07

odds ratio for LV for benefit of trained over untrained talkers at post-test: 4.88/3.47= 1.4

odds ratio for LV over HV benefit for benefit of trained over untrained talkers: 1.43/1.07 = **1.31**

[3a]extracted from beta value for main effect of variability condition from Dong et. 2019 (Note this is most similar to the current study in that in the HV the talkers were organized into blocks during training in a similar manner to the current study)

[3b] extracted from beta value for interaction between test session for one talker versus multi-talker training in Dong et. 2019

[4] Computed from means from Logan, et al. (1991) and Lively, et al. (1993) as follows:

odds of choosing correct option at pre-test across experiments: 2.76

odds of choosing correct option at post-test across experiments: 4.33

odds ratio across conditions for change from pre to post-tests: 4.33/2.76 = **1.57**

[5] odds of choosing correct option in trained talker post-tests across experiments: 4.05

odds of choosing correct option in untrained talker post-tests across experiments : 3.45

odds ratio for benefit of trained over untrained talkers at post-test across experiments: 4.05/3.45 = **1.17**

Results

All data and analyses are available on the OSF page of this project https://osf.io/4hkqb/.

Approved deviation from the methods and analysis described above

Following the acceptance of methods as part of the Stage 1 report, another methodological deviation between ours and the original studies became apparent: When providing feedback during training, the original studies only provide feedback on incorrect trials. We instead provide feedback on *both* correct and incorrect trials. Providing feedback on both correct and incorrect trials is in line with more recent phonetic training studies (e.g. Iverson et al., 2005; Perrachione et al., 2011; Shinohara & Iverson 2021; Giannakopoulou et al., 2017; Dong et al., 2019; Sadakata & McQueen 2013, 2014). Since feedback includes an additional auditory presentation of the stimulus, providing feedback on only incorrect trials could introduce a confound where the amount of incorrect responses is related to the amount of exposure to the stimuli learners receive. Since we predicted that LV learners should be overall more accurate during learning, incorrect-only feedback would mean that HV learners would receive systematically more exposure to the phonetic contrast of interest. This confound can, however, be avoided by giving feedback on both correct and incorrect trials, such that learners receive the same amount of input for both HV and LV training.

Another approved deviation from the methods above arose due to technical issues during data collection: 67 people initially completed the experiment. We, however, noticed an error in the experimental set up that resulted in inconsistencies in how feedback was provided during training. This error is described in more detail on the OSF page of this project (https://osf.io/9fuyx/). All of these 67 participants were replaced and their data was not included.

Participant sample

After the feedback issue was fixed, 212 participants were then recruited. From this sample, we excluded 46 participants in total: 9 participants never started the experiment after initial

recruitment, 4 were rejected at the headphone screening, 3 did not finish the full pre-test, 9 were rejected because they failed the attention checks/catch trials in the first training session, 1 did not complete the full fifteen training sessions, one was excluded due to a technical error, and an additional 19 were excluded from analysis because they scored above 85% on the pre-test.

Our final sample thus included 166 participants, which exceeds our target of 160 participants. The additional 6 had already started the experiment when it became clear the recruitment target had been reached and were thus included in the analysis. As planned, we arrived at the full sample after implementing a Bayesian optional stopping approach: This means that during recruitment, we periodically checked the Bayes Factor of our key effect of interest, which is the effect of variability condition on the amount of change observed from pre-test to novel post-test. Checks were run at 60 participants, 75 participants, 118 participants, 163 participants, and 166 participants, all after excluding the initial 67 participants who received feedback errors. The Bayes Factor for our key test of the hypothesis, i.e. the effect of variability condition on the amount of change observed from pre-test to post-test with novel talkers did not show substantial evidence in either direction during these checks, so recruitment continued.

Of our final sample of 166 participants, 87 identified as female, 77 identified as male, and 2 did not disclose their gender. Participants were on average 20.4 years old (SD = 2.3 years)[7], all but one currently resided in Japan, and all had started learning English either in primary or secondary school, with 94 out of 166 currently taking English lessons. On average, participants had just over 10 years of English experience, ranging from 5 to 21 years of exposure. Participants rated their overall English proficiency on average at 2.86 out of 5 (SD = 0.72) and reported to not speak any English at home. Out of the total sample, 86

[7] Note that due to a technical error in Gorilla, the age data for 17 participants was not recorded, so the age reported here is the mean and standard deviation for 149 participants.

participants were randomly assigned to HV and 80 to LV training condition. At pre-test, participants' initial performance was matched across variability conditions, such that participants in the HV condition gave correct responses on 61.9% (SD = 7.3%) of trials and those in the LV condition gave correct responses for 60.5% (SD = 8.3%) of trials.

Training results

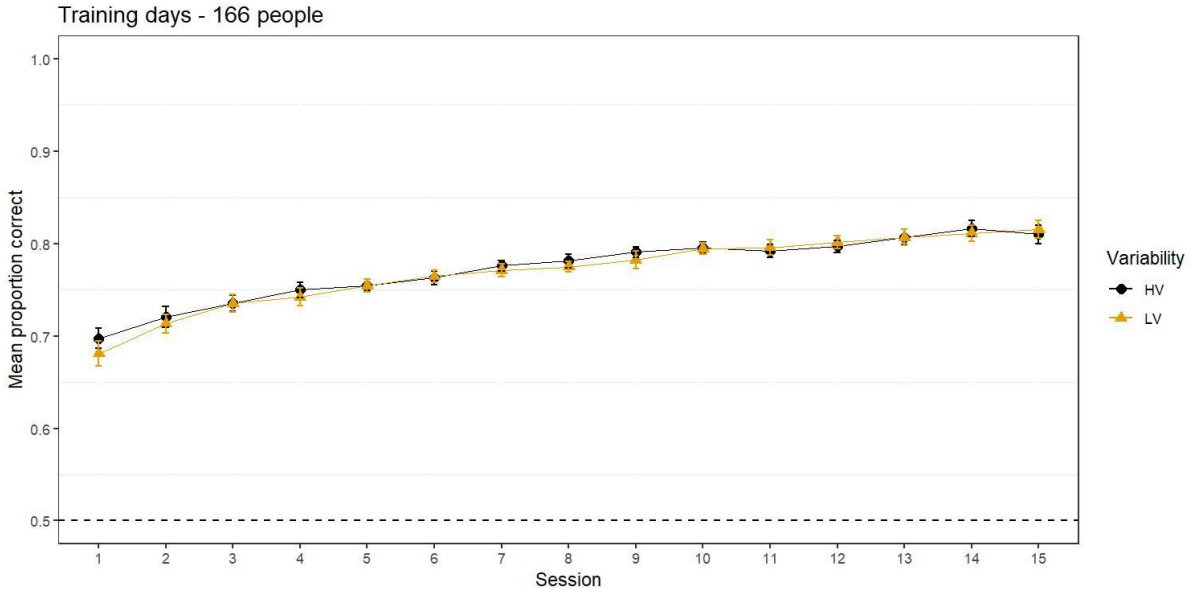


Figure 2. Response accuracy during Training, measured in proportion correct. Shape and colour indicate the variability condition participants were assigned to. Error bars indicate 95% CI; the dashed line indicates chance.

In the first session the average accuracy for the learners receiving HV training was 69.8% (SD = 4.9%) and for the learners receiving LV training it was 68.1% (SD = 5.9%). Both groups showed clear improvement across the fifteen sessions suggesting an effect of training: by the last training session, the HV learners had improved to 81.0% (SD= 4.8%), and LV had improved to 81.5% (SD = 4.3%). This means that HV learners showed an improvement of 11% and LV learners improved 13%, which is in line with the usual 10-15% improvement seen in most phonetic training studies (see Figure 2).

We ran a logistic mixed effects model to test whether variability affected listeners' performance during training. The final model structure was the following: *accuracy ~ session*Variability + (session|participant) + (1|target)*. The models converged with the Nelder-Mead optimizer, after all attempts at convergence with BOBYQA (as specified in the methods above) failed. As described above, we extracted the coefficients for the fixed effects corresponding to our hypotheses and tested the evidence by applying Bayes factors.

The evidence for the interaction of variability condition by test session, in the direction of more improvement for the LV learners, is substantial for the null (β = 0.006, SE = 0.008, z = 0.696, p = .487; $BF_{H(0,0.8)}$ = 0.022 [RR 0.115, >5.515]). This indicates that contrary to the hypothesis, the LV learners did not improve more in their ability to discriminate between /r/ and /l/ across training. While the choice of prior for Bayesian analyses can affect results, it is unlikely that it did so here: the robustness regions indicate that we would have found evidence for the null if we had used a prior that was 1/8th the size of our current one and thus much smaller (as well as any prior larger than the one we assumed, since larger priors bias finding evidence for the null).

Since it can be harder to get evidence for an interaction than a main effect, we also planned to look at whether overall during training the LV learners outperformed the HV learners. The accuracy data from the first session is not meaningful here since both training groups are trained on a single talker per session, such that there is no difference between HV and LV input during the first training session. For this reason, we also ran an analysis looking at the main effect of variability for all sessions from second training session onwards. The model structure was identical to the one reproduced above, and the model also converged with the Nelder-Mead optimizer. In this analysis, we found substantial evidence for the null in the main effect of variability condition (β = -0.017, SE = 0.134, z = -0.124, p = .901; $BF_{H(0,0.35)}$ = 0.326 [RR 0.415, >5.515]). This means that from the point at which the two variability conditions are differentiated in the training paradigm (i.e. session 2, when a second, novel talker is presented for HV learners), the evidence suggests that the LV learners did not

perform any better than HV learners. We acknowledge here that using a smaller prior would have led to the evidence being evaluated as ambiguous, although we would continue to find substantial evidence for the null for any larger prior.

Model 1: Key test of hypothesis - pre-test to novel post-test

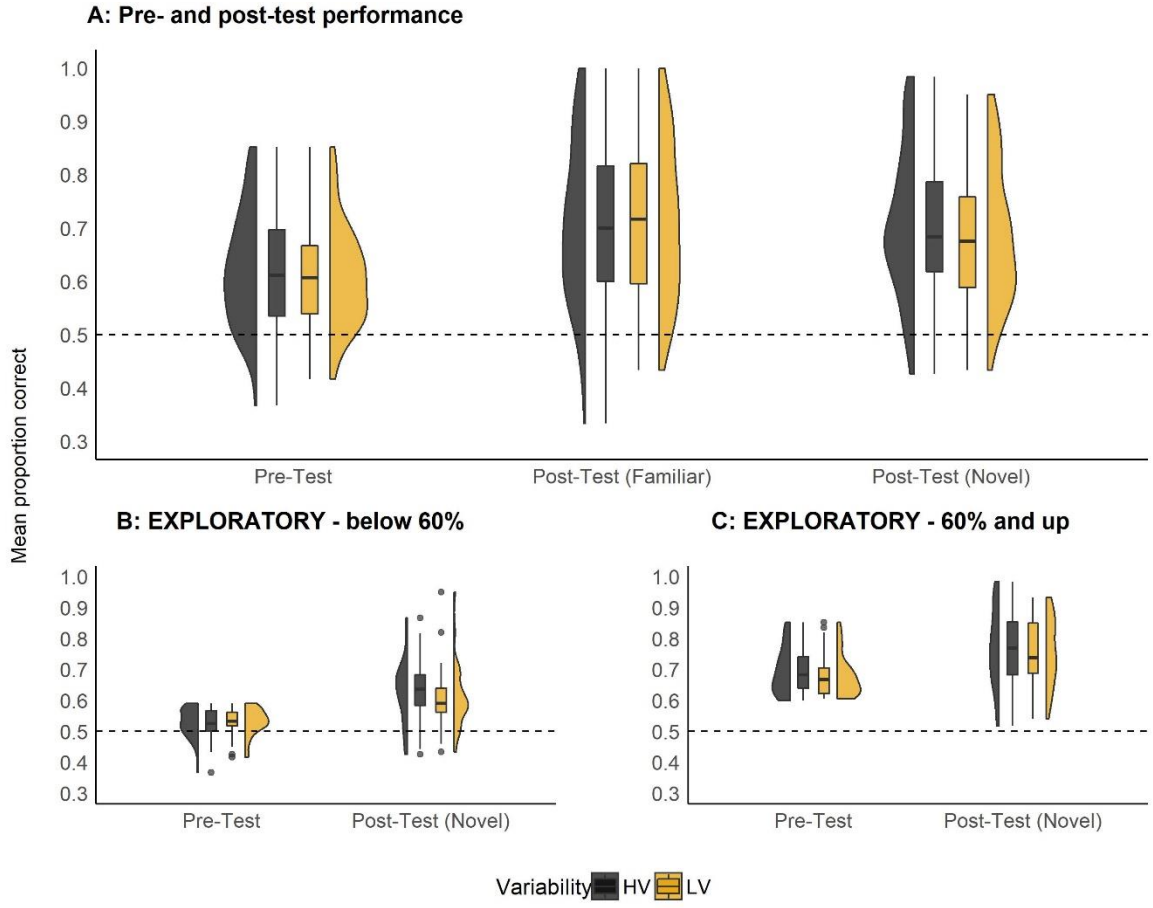


Figure 3. This figure demonstrates the overall pattern of performance across pre- and post-tests, as well as the exploratory analyses relating to our key hypothesis. Graph A shows half-violin plots and boxplots visualising performance (measured in mean proportion correct per participant) on pre-test and the two generalisation tasks at post-test, with familiar and novel talkers. Graph B and C show half-violin plots and boxplots visualising the exploratory analyses into pre-test aptitude differences: Graph B presents pre-test and novel talker post-test scores for participants who scored below 60% accuracy at pre-test. Graph C shows pre-test and novel talker post-test scores for participants who scored 60% or above at pre-test.

Note that the y-axis starts from 30% accuracy in all panels. The dashed line in all panels indicates chance performance.

At pre-test, HV learners were on average 61.9% correct (SD = 7.3%), while LV learners reached 60.5% accuracy (SD = 8.3%). In the novel post-test, where all participants completed the /r/-/l/ discrimination based on a previously unheard talker, HV learners were on average 70.4% correct (SD = 6.5%) and accuracy for LV learners was 68.1% (SD = 6.3%). To test our key analysis – whether variability affected listeners' performance after training – we fit another set of logistic mixed effect models with the structure: *accuracy ~ Test*Variability + speaker + (Test|participant) + (1|target)*[8]*.* We extracted the relevant coefficients to test using Bayes factors.

We first establish whether HV and LV learners each show improvement from pre- to post-test. Looking at simple effects establishes that this is the case: there is substantial (indeed very strong) evidence of pre- to post- improvement in both HV learners (β = 0.448, SE = 0.085, z = 5.298, p < .001; $BF_{H(0,0.45)}$ = 284830.477 [RR 0, >4.067]) and LV learners (β = 0.372, SE = 0.083, z = 4.492, p < .001; $BF_{H(0,0.45)}$ = 6266.237 [RR 0, >4.067]). This means both groups show gains following training (in line with the evidence of improvements across session in the training data) and here show generalisation to novel talkers. The robustness regions indicate that we would have found substantial evidence for learning no matter what prior we used in both of these models.

Our key effect of interest is to examine whether there is a difference in the amount of improvement from Pre-test to Generalisation 1 (with novel talker) depending on amount of

[8] Fixed and random effects for Test and Speaker were centred across all models in the analyses reported here, to avoid multicollinearity. Variability was centred across models as well, except where models were run to specifically test for simple effects, i.e. testing to see if there was a main effect of Test in a model containing *only* data from the HV condition, or *only* data from the LV condition. In those cases, the models were re-run without centring the effect of Variability. Instead, they had an uncentred effect of Variability, which had either the HV condition set at the reference level for the effect of Test (for the model investigating the effect in HV condition only), or LV condition set as reference level. This way, the results from the simple effect of Test in each of those models will show whether there is evidence for change across tests in each condition.

variability experienced during training. As can be seen in Figure 3, there is some numerical difference between the two variability conditions with HV learners showing more learning than LV learners (8.5% in HV learners versus 7.6% in LV learners respectively). However, statistically the results are less clear: the evidence for the interaction between test session and variability condition was ambiguous (β = 0.076, SE = 0.090, z = 0.844, p = .398; $BF_{H(0,0.32)}$ = 0.598 [RR 0, 0.567]). The robustness regions indicate that there is no smaller value we could have used which would have led to an evaluation of substantial evidence for a benefit of high variability; to obtain evidence for the null, we would have had to use a prior almost twice that of our current prior.

Model 2: novel post vs familiar post

We then ran a logistic mixed effects model to compare performance on the two post-test tasks, comparing items spoken by a novel talker (Generalisation 1) or familiar talker (Generalisation 2). The final model structure was: *accuracy ~ Test*Variability + speaker + (Test|participant) + (1|target)*. Coefficients for key fixed effects were extracted for Bayes Factor analyses. In Generalisation 1 (novel talker), HV learners achieved 70.4% accuracy (SD = 6.5%) while LV learners performed at 68.1% accuracy (SD = 6.3%). In Generalisation 2 (familiar talker), accuracy for HV learners was 70.8% (SD = 6.7%), and 71.0% for LV learners (SD = 7.5%). In this analysis (following Lively et al., 1993) we are looking for a greater difference between the two post-test tests for the LV learners than the HV learners, whereby the LV learners show higher accuracy with the trained talker and HV learners show no difference or a weaker effect. The means are in line with this prediction and we find substantial evidence for the null for an effect of test session for the HV learners (β = 0.049, SE = 0.133, z = 0.367, p = .713; $BF_{H(0,1.17)}$ = 0.154 [RR 0.567, >4.067]), However for the LV learners, the evidence for this comparison is ambiguous (β = 0.173, SE = 0.134, z = 1.293, p = .196; $BF_{H(0,1.17)}$ = 0.464 [RR 0, 1.567]). This indicates that HV learners generalised and did not treat stimuli spoken by novel or familiar talkers differently, while for LV learners this evidence is not conclusive. The robustness regions indicate that we would have found the

same result for HV learners for a prior that was half the size of our current one (and any priors larger than the one we assumed); for LV learners, there is no smaller prior which would have led to our finding evidence for a difference between conditions, however of prior 1 1/3rd times larger would have led to finding evidence for the null.

Turning to the interaction, which provides the key test of the hypothesis, we similarly found ambiguous evidence for the interaction which would indicate a greater benefit of familiarity for LV learners ($\beta = 0.124$, SE = 0.083, z = 1.496, p = .135; $BF_{H(0,0.27)} = 1.504$ [RR 0, 1.367]). This means that results are inconclusive as to whether the amount of variability encountered during training results in differential performance for novel and familiar talkers on generalisation. The robustness regions indicate that there is no prior that would have led to us finding evidence for the difference, and we would only have found evidence for the null using a prior that was over four times the size of our current one.

Exploratory Analyses

Thus far we have not found substantial evidence for a benefit of high-variability training over low-variability training. In the introduction, we noted that some previous work had found that rather than seeing a group-level benefit of high-variability training, there may be a high-variability benefit for a subset of higher aptitude learners, and maybe even a low-variability benefit for lower aptitude learners (Sadakata & McQueen, 2014; Perrachione et al., 2011). We also noted that testing this claim with adequate power would be extremely difficult due to the effect of interest being a three-way interaction between Test, Variability and Aptitude, and thus we did not include this in our pre-registered analyses. Nevertheless, we did perform several exploratory analyses looking at whether there is evidence for different variability effects in different subgroups and at the relationship between condition differences and various individual difference measures included in the study. The goal is to see if these can shed light on why we find ambiguous evidence for a variability effect in the main analyses reported above.

Our exploratory analyses were twofold[9]: 1) we investigate the role of pre-test "aptitude" by looking at whether learner groups split by higher vs lower pre-test accuracy might separately benefit from either HV or LV input, and 2) we investigate further links with individual difference measures of attention, language ability and beyond. All data and scripts for these exploratory analyses are available on the OSF page of this project. We stress again that none of the analyses reported in the sections below were planned or pre-registered, so the reported results need to be interpreted with this in mind.

Pre-test aptitude

The following analyses investigate whether the lack of evidence for a high variability advantage in our data might be due to a difference between participants who had a lower or higher initial "aptitude" in distinguishing /r/ and /l/. "Aptitude" in the current analysis refers to the performance at pre-test. Based on findings from studies of lexical tone training (Sadakata & McQueen, 2014; Perrachione et al., 2011 but see Dong et al., 2019; Zhang et al., 2018), we tentatively predicted that participants who performed better at pre-test might benefit more from HV input, while those who performed worse at pre-test might benefit more from LV input. If this were the case in our sample, these opposite patterns of improvement could have cancelled out any overall benefit in our main analysis.

To investigate such possible aptitude effects, we ran the main analysis which is the key test of our hypothesis – i.e. looking for difference in increases from pre-test to novel post-test (Model 1 above) - splitting out participants into two groups: a group that performed above

[9] In addition to the analyses reported here, we conducted an additional analysis where we repeated the main analyses above but included the data from the participants for whom there were intermittent issues with the feedback during training (as reported in the 'Approved deviations' section above) to see if a bigger sample might change our findings. Sixty of the excluded participants would have met the inclusion criteria if it were not for the feedback issue, resulting in an overall total of 226 participants: 117 completed HV training, and 108 completed LV training. All model structures were identical to the ones described for the relevant analyses above. Full analyses are reported on the OSF but in sum we obtained a similar pattern of results and in particular, even with these additional participants the evidence for a benefit of variability on generalization remains ambiguous across the different tests.

chance at pre-test, and a group that performed at or below chance at pre-test. For this purpose, we calculated the 95% confidence intervals around chance performance in our pre-test task. This was achieved by running 100,000 permutations of random responses within our tasks (60 trials for a two-way forced choice paradigm). This simulation indicated that the upper bound of the 95% CI was 60%. We therefore divided the learners into two groups: those who scored below 60% accuracy at pre-test (and therefore could be considered to be indistinguishable from or even below chance), and those who scored 60% or more at pre-test. Seventy-seven people had a pre-test score below 60% (39 LV, 38 HV), while 89 people had a pre-test score of 60% and above (41 LV, 48 HV).

With this subdivided sample, we performed two analyses: 1) We asked whether there is evidence for an HV benefit on pre- to post-test improvement in the group that has pre-test scores of 60% or higher. 2) We asked whether there is evidence for an HV benefit *or* evidence for an LV benefit in the group that has pre-test scores below 60%. Note the latter analysis tests two separate directional hypotheses: An HV benefit in the lower aptitude pre-test group would be in line with our overall hypothesis, while an LV benefit in the lower aptitude group would be in line with some of the previous literature discussed above. In terms of priors, for the Bayes Factor analyses for HV benefit, in each case we can use the same prior as for the pre-registered analysis. For LV benefit it is unclear what value to use. Ideally the prior would be based on estimates of the actual effect sizes in previous studies that found a benefit of LV training in one of the groups. However, given the mixed results in the literature and differences in paradigms used, we found it difficult to determine what study to base any prior off. Thus, we decided to use the same value as for the HV benefit (i.e. guessing that if there is such a benefit, it will be around the same size as that found for HV in the original studies), although testing for an effect in the opposite direction. Critically, as in all our other analyses we report a robustness region, which allows readers to evaluate how much results would have changed had we used a different prior.

The overall pattern of change can be seen in Figures 3 B and C. The group starting at *60% accuracy or above*, in the HV condition improved 7.13%: they were accurate on 69.74% (SD = 7.4%) of trials at pre-test and improved to 76.87% correct (SD = 12.8%) at the post-test with novel talkers. The LV condition of this same group similarly improved 7.63%: they started at pre-test at 67.97% (SD = 6.7%) and improved to 75.60% (SD = 10.8%) in the novel talker post-test. Numerically, those high aptitude learners in the LV condition thus improved 0.5% more than those in the HV condition. Using the same model formula as in Model 1 reported above and testing for a session by variability-condition interaction representing greater improvement in the HV condition, we found substantial evidence for the null (β = 0.006, SE = 0.078, z = 0.074, p = .941; $BF_{H(0,0.32)}$ = 0.254 [RR 0.2667, 4.0667]), i.e., there was evidence against there being an HV benefit in this high-aptitude group.

In the group starting *below 60% accuracy*, learners in the HV condition improved by 10.1%: they had a pre-test accuracy of 52.66% (SD = 4.9%) and improved to 62.76% (SD = 9.7%) in the novel talker post-test. Those in the LV condition improved by 7.53%: they had a pre-test accuracy of 53.05% (SD = 4.3%) and improved to 60.58% (SD = 9.7%) in the novel talker post-test. Numerically, the low aptitude learners in the HV condition thus improved 2.6% more than those in the LV condition. Note that a model identical to Model 1 in the planned analyses above did not converge so correlations between by-participant slopes were removed to allow convergence. As laid out above, we checked *two* directional hypotheses: checking for an interaction in the direction of more improvement for the LV condition (LV benefit), and separately checking for an interaction in the direction of more improvement for the HV condition (HV benefit). We found substantial evidence *against* a LV benefit (β = -0.059, SE = 0.054, z = 1.089, p = .276; $BF_{H(0,0.32)}$ = 0.086 [RR 0.1667, >4.067]), while evidence for an HV benefit was ambiguous (β = 0.059, SE = 0.054, z = 1.089, p = .276; $BF_{H(0,0.32)}$ = 0.512 [RR 0, 0.4667]).

Overall, none of these analyses showed evidence for aptitude-based differences in training variability benefit: we find evidence against the hypothesis that those with higher aptitude at

pre-test improve more in HV. Similarly, we find evidence against the hypothesis that learners with lower aptitude at pre-test improve more in LV. Thus, we do not find evidence that differences in aptitude - at least as measured by pre-test score – prevented us from seeing an overall HV benefit in our main analysis.

Individual differences

Our final set of analyses explored links between individual difference measures and potential differences in benefits from variability input. The individual difference measures for which we collected data are described in more detail in Table 2 above, and included measures of attention, auditory processing and cognitive control. Here, since there is no clear method for splitting participants into "high" and "low" aptitude for these measures, we instead take a continuous approach and hypothesized that better performance on these measures might help learners benefit from high variability training, leading to greater pre- to post-test improvement. For each of our individual difference measures, we tested for a three-way interaction between the individual difference measure, test-session, and variability-condition. We do not attempt to compute Bayes Factors here since it is unclear what values to base the prior on.

Separate exploratory models with interactions between the individual difference measure and the main factors of interest (variability condition and pre/post-test session) were run for each individual difference measure[10]. Relevant here, however, is that a significant three-way interaction with test-session and condition was only found in only one of the measures (not correcting for multiple comparisons): the AXCPT response time measure (β = -0.080, SE = 0.0325, z = -2.473, p = .013). This task has been used to measure constructs such as cognitive control (Gonthier et al., 2016) and sustained attention (Halperin et al.1991; Rosvold

[10] We also looked to see which measures were predictive of (i) overall performance (significant main effect) (ii) were predictive of pre- to post-test improvement (involved in a significant two-way interaction with test-session). Results were: (i) melody memory (ii) melody memory, dichotic listening, and AXCPT response time.

et al., 1965), although there are many ways that performance on the task can be measured as well as debate about their interpretation (Riccio et al, 2002). Here we found that overall slower response time was associated with more gains in the HV condition, in other words, the interaction is in the opposite direction than would be expected if more cognitively able leaners were more able to benefit from HV input. Given this, and the lack of significant three-way interaction with any other measure, this exploratory analysis finds no evidence that higher aptitude participants benefit more from HV training. However, given that we do not use Bayes Factors here, we cannot draw strong conclusions that aptitude plays no role in modulating variability benefits. More detailed analyses to qualify any variation due to individual differences are beyond the scope of this paper to do such analyses justice.

Discussion

Our study set out to replicate seminal phonetic training studies investigating the effect of talker variability on generalisation ability (Lively et al., 1993; Logan et al., 1991) in a large participant sample. To this end, 166 Japanese learners of English participated in our experiment. Learners completed a pre-test to determine their existing ability to distinguish between the non-native phonetic contrast /r/ and /l/, followed by 15 sessions of either HV (5 talkers) or LV (1 talker) phonetic training. After training, learners then completed two generalization post-tests (novel talkers and/or novel items) to establish the efficiency of the HV vs the LV phonetic training. The key research question we aimed to answer was whether learners are able to *generalise* more after having had phonetic training on multiple talkers. Bayes factor analyses reveal very strong evidence that participants in *both* conditions improved in their ability to distinguish between /r/ and /l/ from pre-test to the post-test with novel talkers (8.5% after HV training and 7.6% after LV training) – indicating successful generalization of learning to new talkers. However, the evidence that these gains were greater after HV training was ambiguous, such that we do *not* find substantial evidence for more generalization following HV than LV training, nor do we find substantial evidence for the null. Following the analysis of Logan et al. (1991) and Lively et al. (1993), we examined

whether LV learners would show a greater difference between a trained and a novel talker in the post-test than did HV learners, which would have also indicated poorer generalisation of learning after LV training. Evidence for this HV benefit was again ambiguous.. Further exploratory analyses suggested that even when only looking at people who performed above chance at pre-test, evidence for an HV benefit in pre- to post improvement remains ambiguous. In sum, in no analysis did we find substantial evidence for a specific benefit of multi-talker training over single talker training. On the other hand, it is important to acknowledge that although we did not find substantial evidence *for* this hypothesis, we also did not find evidence for the null. As our sample size analyses showed, in this training paradigm, obtaining evidence for the null when it is true is extremely hard with realistic sample sizes (see Figure 2). Furthermore, our sample size of 166 participants is already considerably larger than the largest sample sizes in the existing literature (N = 60, Dong et al., 2019; N = 64, Perrachione et al., 2011) with our power calculation indicating that we had a 99% chance of detecting a HV benefit in our key analysis if it was truly there. In this context, the ambiguous Bayes factors then indicate to us that if an effect exists, it is likely much smaller than the effect reported by Logan et al. (1991) and Lively et al. (1993).

HV benefits in phonetic training: Contextualising our findings in the existing literature

Overall, although the findings of this replication are different from those of the original study, they are broadly consistent with the extensive literature review in the introduction, which found relatively few studies that compared HV vs LV training directly, and mixed results for those few that did. We note, however, that the view that recent findings are mixed is not held by all researchers in the field: For example, Zhang et al. (2021) have published a new systematic review and meta-analysis of HV benefits in phonetic training, which found evidence of an HV benefit for generalization, with a large effect size. Although a full critique of the meta-analysis is beyond the scope of the current paper, the overall different finding is in part explained by the fact that there is only limited overlap between the literature reviewed in our literature review and the 10 experiments included in their meta-analysis. This is due to

some studies included by Zhang et al. (2021) being unpublished PhD theses or very recent papers post-dating our literature search. More surprisingly, there are studies which we included in our review which were excluded from Zhang et al. (2021)‘s generalisation analysis, despite apparently meeting the criteria for inclusion in the meta-analysis (e.g. Giannakopoulou et al. [2017] and Dong et al. [2019]), both studies that do not report a HV benefit. While this recent meta-analysis illustrates that the existing evidence base for a HV benefit in phonetic training can be interpreted in different ways, the current replication study finds no substantial evidence for a HV benefit. In the followings sections, we will therefore discuss how these findings may fit within the existing literature, focussing on possible explanations.

Our results clearly confirm the seminal finding and subsequent replications that the phonetic training paradigm is useful and a means of improving learners abilities to identify non-native phoneme contrasts. However, not finding clear evidence to show that exposure to multiple talkers improves learning, and specifically generalization of learning, over and above single talker training is surprising. Theoretically, we expect variability over irrelevant talker-specific cues to promote identification and generalization of *relevant* phonetic cues.

It is worth noting that a recent study – published since our review - found a parallel result while attempting to replicate a related finding by Bradlow and Bent (2008), who found a multi-talker advantage for training listeners to recognize foreign-accented speech and generalize to novel talkers. Xie et al. (2021) performed a high-powered replication of this study and found that learning and generalization improved with both single-talker and multi-talker training, while finding limited evidence that the two were different. Thus, both of these replication studies came to the same conclusion: multi-talker training might not confer a specific advantage. Interestingly, like the current study, Xie and colleagues used more than one talker for the single talker training condition (counterbalanced across participants), and found that some of the talkers seemed to give less of a training benefit than others. This

suggests the possibility that the replicability of the multi-talker advantage might depend on chance differences in choice of talker for the single-talker condition.

Different types of variability: When is high variability training more beneficial than low variability training?

Another possibility is that finding strong generalisation of learning following both HV and LV training could be due to the number of other types of variability that are present in the stimuli. In particular, in the current replication, and the original studies, HV and LV learners heard 68 minimal pairs (136 words) exhibiting the contrasts in a variety of phonetic environments. It is possible that the amount of variability contained in these stimuli is already sufficient to boost detection of the relevant cues in a way that also boosts generalization to novel talkers.

The amount of variability in the stimuli in our experiment may explain another unexpected finding: Based on the previous literature, we predicted that during training itself, LV learners would find the task easier. This was, however, not the case in our study with LV learners showing the same learning trajectories as HV learners. Although Logan and colleagues (1991) did not report on this aspect of their data, some literature has found an LV benefit during training itself (Brekelmans et al., 2020; Dong et al., 2019; Evans & Martín-Alvarez, 2016; Giannakopoulou et al., 2017; Perrachione et al., 2011). However, each of these studies included far fewer items in training, using between 18 words and 80 words compared to the 136 used in the current study. There are other differences between our training task and these other studies; however, we believe that the LV benefit in those experiments at least partially reflects the fact that their participants were able to quickly familiarize themselves with a relatively small set of specific stimuli, which was not the case in our experiment.

This then brings us to an important open question: what is it about multiple talker variability that is supposed to be helpful, and how much of this variability is needed to get a benefit? The fact that we find robust improvement and generalization in the LV condition in the current study provides a positive contribution to the phonetic training field, as it suggests

a single talker might already provide enough variability to allow learners to generalise with similar success for the LV and HV training. This can lead to further questions as to what kind of variability is being manipulated and how, why, and to what extent the variability manipulation may facilitating learning. These questions align neatly with recent work by Raviv et al., (2022), who suggest that studies which investigate “variability” across different fields do not share a single coherent concept of what variability is and how it may benefit learning. Instead, they can be understood in terms of four underlying types of variability: numerosity (differences in set size: few vs many stimuli), heterogeneity (differences in stimulus similarity: many relatively similar stimuli vs many relatively less similar stimuli), contextual diversity (differences in context: stimuli are embedded in a similar context vs in diverse contexts), and diversity of training schedule (differences in presentation/schedule of training: blocked vs interleaved). The talker variability effects within the phonetic training paradigm are primarily examples of variability in heterogeneity where listeners are all trained on the same number of stimuli, but the stimuli should be less similar to one another in the HV training condition due to being sampled from different talkers. However, from a different perspective, phonetic training studies could be (and have often been) interpreted as manipulating variability via changes numerosity, when regarding the number of talkers (single vs multiple) as the most perceptually meaningful unit to quantify variability. Of course, in light of not finding substantial evidence for an HV benefit in our study, it could then in the same vein be argued that the single vs multiple talker manipulations used in phonetic training paradigms do not manipulate high versus low variability at all, at least not in a perceptually meaningful way: If we were, for example, to believe that only contextual variability would matter, both LV and HV training paradigms include the same amount of contextual variability in our and other studies. This attempt of trying to classify the type of variability that has (or has not) been manipulated in phonetic training studies highlights that variability manipulations have perhaps been implemented in a somewhat simplistic manner. Variability manipulations are often implemented as unitary concepts, instead of viewing the standard

variability manipulation as highly multivariate, affecting different types and sources of variability that may or may not all be meaningful during learning.

This is, of course, understandable as, historically, within the field of phonetic training the inter-talker variability effects initially showed compelling and fruitful findings. As a result, however, there is relatively little research investigating the potential benefits of other different *types* of variability because multiple talker variability was quickly accepted as the 'gold standard' in the field. Most studies that do investigate different types of variability focus on the traditional HV conditions only, with adjustments tend to be made to this condition with the 'default' multiple-talker condition serving as a baseline comparison. Examples of this research include studies investigating talker blocking versus intermixing (manipulating what Raviv et al. 2022 call 'training schedule', e.g. Dong et al., 2019; Zhang et al., 2021), adding explicit instruction (Wiener et al., 2020), and adding a visual modality (Hardison, 2003). This apparent gap in the literature leaves an exciting avenue of research to further uncover what facets of variability may facilitate the learning of novel phoneme contrasts.

High variability benefits and individual differences in participants

In addition to HV benefits being dependent on the types of variability manipulation, another suggestion in the literature is that the ability to benefit from variability might depend on participant factors. Previous studies raised the possibility that some measure of aptitude could modulate training effects: HV training is thought to include more useful variability than LV training. At the same time, however, HV training is also thought to at least initially present more perceptual challenges due to its complexity. As a result, lower aptitude learners might struggle to cope with HV training, thus failing to access or perceive the useful variability in a meaningful way. In the current data, however, our exploratory analyses found no evidence that either aptitude (defined as accuracy at pre-test), nor performance on various tasks assessing measures of individual differences (such as auditory processing and attention) was linked to how HV vs LV input affect training outcomes. However, since these analyses were exploratory, dedicated experiments would be needed to test for these effects in a

targeted manner. Since our results seem to suggest that a HV benefit is either smaller than originally reported in the studies or that it may interact with participant properties in complex ways, the main challenge for such dedicated studies is to achieve adequate power.

To achieve this, one approach might be to simply power and test a three-way interaction between pre-test aptitude, generalization, and training condition. As pointed out in the introduction, this would be a methodological challenge given the sample size that such a study would likely need. We therefore believe that an important first step is to first establish the relationship between various aptitude measures and the effectiveness of phonetic training for a single training condition in isolation. This is easier to statistically power as it would only require a 2-way interaction, and potentially allows a method for building a theoretically, or at least empirically, motivated account of "aptitude" in these paradigms. Where strong predictor(s) of learning are found, we can then consider whether participants who are weaker in these abilities might react differently to input with different levels of variability. This could provide informative priors on effect sizes which could be used to design a study testing how the aptitude measurement affects variability benefits with sufficient statistical power.

Conclusion

The current study attempted to replicate the findings of seminal work by Lively and colleagues (1991) and Logan and colleagues (1993), which found that phonetic training with multiple talkers led to greater generalization to novel talkers than single talker training. Across two different measures, our study found no evidence of such effect in a sample of 166 participants. Bayes Factor analyses found that in each case the evidence for a variability benefit was ambiguous. This means that null effects are not established but also indicates that this type of training paradigm is not sufficiently sensitive to detect variability benefits. We suggest that, if it does exist, such a benefit in phonetic training is likely very small. On the other hand, our results here clearly shown that even training with a single talker can lead to robust improvements in generalization, at least when the items in training are sufficient

varied, as was true here and in the original studies. Our findings thus overall add to a literature that often reports conflicting findings. In light of these conflicting findings, we believe that our study thus helps highlight that HV benefits in phonetic training are unlikely monolithic effects.

HV benefits are, however, well-attested far beyond phonetic training, with HV benefits being reported for language learning, visual category learning, as well as motor learning (Raviv et al., 2022). Beyond empirical work, HV benefits are also predicted by mechanistic accounts of learning across fields and are supported by computational modelling. In light of this, it is therefore likely that variability is indeed beneficial for phonetic training. However, it remains unclear which aspects of the variability experienced during training (e.g., talker vs stimulus variability) may be informative and how much of such variability is necessary to facilitate later generalisation of learning for phonetic training. There is therefore a clear case for future work to determine how and under which circumstances variability can support and boost the efficacy of phonetic training. With an increasingly detailed evidence base outlining the role of variability during phoneme learning, a mechanistic account of the learning process can be built and integrated into theoretical models of learning in the context of speech and beyond.

Acknowledgements

This research is funded by a British Academy/Leverhulme Small Research Grant (SRG2021_210374) awarded to Gwen Brekelmans, an NSERC Discovery Grant (NSERC RGPIN-2021-04117) awarded to Meghan Clayards, and a St Johns College Early Career Researcher grant awarded to Elizabeth Wonnacott. Nadine Lavan is sponsored by a Sir Henry Wellcome Fellowship (220448/Z/20/Z). The funders had no role in the study design, data collection, analysis, and interpretation, preparation of the manuscript, or decision to submit for publication. We would like to thank John Logan, Ann Bradlow, and David Pisoni for sharing the original stimuli lists, Matthew Jaquiery for help with programming the

experiment, Anastasia Giannakopoulou for making available a dataset that we used in our sample size simulation, Axelle Calcus, Adam Tierney, and Kazuya Saito for sharing tasks we used in the individual differences tests, and Yasuaki Shinohara (Waseda University), Michael Honywood (Shinshu University), Greg Hadley (Niigata University), Keiko Hanzawa (Tokyo University of Science/Waseda University), Risa Nabei (Tokai University), Kozue Nakatsuka (Reitaku University), and Takayuki Nagamine for help in recruiting participants.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods.* https://doi.org/https://doi.org/10.1101/438242

Antoniou, M., & Wong, P. C. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *The Journal of the Acoustical Society of America*, *138*(2), 571-574. https://doi.org/10.1121/1.4923362

Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, *35*(6), 1105-1138. https://doi.org/10.1111/j.1551-6709.2011.01181.x

Atkinson, M., Kirby, S., Smith, K. (2015). Speaker Input Variability Does Not Explain Why Larger Populations Have Simpler Languages. *PLoS ONE 10*(6): e0129463. https://doi.org/10.1371/journal.pone.0129463

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, *59*(4), 390-412. https://doi.org/10.1016/j.jml.2007.12.005

Baker, K. A., Laurence, S., & Mondloch, C. J. (2017). How does a newly encountered face become familiar? The effect of within-person variability on adults' and children's perception of identity. *Cognition*, *161*, 19-30. https://doi.org/10.1016/j.cognition.2016.12.012

Barcroft, J., Sommers, M. S., Tye-Murray, N., Mauzé, E., Schroy, C., & Spehar, B. (2011). Tailoring auditory training to patient needs with single and multiple talkers: Transfer-appropriate gains on a four-choice discrimination test. *International Journal of Audiology*, *50*(11), 802-808. https://doi.org/10.3109/14992027.2011.599868

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387-414. https://doi.org/10.1017/S0272263105050175

Barcroft, J., & Sommers, M. (2014). EFFECTS OF VARIABILITY IN FUNDAMENTAL FREQUENCY ON L2 VOCABULARY LEARNING: A Comparison between Learners Who Do and Do Not Speak a Tone Language. *Studies in Second Language Acquisition, 36*(3), 423-449. https://doi.org/10.1017/S0272263113000582

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boersma, P. & Weenink, D. (2015). *Praat: doing phonetics by computer* [Computer program]. Retrieved from http://www.praat.org/

Bolker, B. (2020). *GLMM FAQ*. Retrieved January 3, 2021, from http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#should-i-treat-factor-xxx-as-fixed-or-random

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l: Long-term retention of learning in

perception and production. *Perception & psychophysics*, *61*(5), 977-985. https://doi.org/10.3758/BF03206911

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

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

Brekelmans, G.**,** Evans, B. G., & Wonnacott, E. (2020). Training child learners on non-native vowel contrasts: the role of talker variability. *PsyArXiv.* https://doi.org/10.31234/osf.io/63dhn

Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, *34*(3), 419. https://doi.org/10.1017/S0142716411000750

Cohen, J.D., Barch, D.M., Carter, C.S., & Servan-Schreiber, D. (1999). Schizophrenic deficits in the processing of context: Converging evidence from three theoretically motivated cognitive tasks. *Journal of Abnormal Psychology, 108,* 120-133. https://doi.org/10.1037//0021-843x.108.1.120

Clopper, C. G., & Pisoni, D. B. (2004). Effects of Talker Variability on Perceptual Learning of Dialects. *Language and Speech*, 47(3), 207–238. https://doi.org/10.1177/00238309040470030101

Davies, M. (2008) *The Corpus of Contemporary American English (COCA).* Available online at https://www.english-corpora.org/coca/.

Deng, Z., Chandrasekaran, B., Wang, S., Wong, P.C.M. (2019). Training-induced brain activation and functional connectivity differentiate multi-talker and single-talker speech training. *Neurobiology of Learning and Memory,* 151. 1-9. https://doi.org/10.1016/j.nlm.2018.03.009

de Heide, R., Grünwald, P.D. (2020). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Rev*iew. https://doi.org/10.3758/s13423-020-01803-x

Dienes, Z. (2008). Understanding psychology as a science: An introduction to scientific and statistical inference. Macmillan International Higher Education.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology,* 5, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural Methods in Consciousness Research*. Oxford: Oxford University Press, pp 199-220.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78-89. https://dx.doi.org/10.1016/j.jmp.2015.10.003

Dienes, Z. (2021). How to use and report Bayesian hypothesis tests. *Psychology of Consciousness: Theory, Research, and Practice, 8*(1), 9-26*.* https://doi.org/10.1037/cns0000258

Dong H., Clayards M., Brown H., Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ* 7:e7191. https://doi.org/10.7717/peerj.7191

Estes, K. G., & Lew-Williams, C. (2015). Listening through voices: Infant statistical word segmentation across multiple speakers. *Developmental psychology*, *51*(11), 1517–1528. https://doi.org/10.1037/a0039725

Evans, B. G., & Martín-Alvarez, L. (2016). Age-related differences in second-language learning?A comparison of high and low variability perceptual training for the acquisition of English /i/-/ɪ/ by Spanish adults and children. *Proceedings of New Sounds.: 8th International Conference on Second Language Speech.* Aarhus, Denmark. Retrieved from https://conferences.au.dk/newsounds2016/

Evans, B.G., Martín-Alvarez, L., & Wonnacott, E. (2017). The effect of variability on phonetic training for adults and children. *Workshop on Speech Perception and Production across the Lifespan*. London, UK. *April 2017*. Retrieved from: http://sppl2017.org/wp-content/uploads/2017/04/SPPL2017_book-of-abstracts.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191. https://doi.org/10.3758/BF03193146

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. Perspectives on Psychological Science, 7(6), 555–561. https://doi.org/10.1177/1745691612459059

Flege, J. & MacKay, I. (2004). Perceiving vowels in a second language. *Studies in Second Language Acquisition,* 26, 1-34. https://doi.org/10.1017/S0272263104026117

Fuhrmeister, P., & Myers, E. B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*, *142*(5), EL448–EL454. https://doi.org/10.1121/1.5009688

Galle, M.E., Apfelbaum, K.S., & McMurray, B. (2015) The Role of Single Talker Acoustic Variation in Early Word Learning. *Language Learning and Development, 11*(1), 66-79. https://doi.org/10.1080/15475441.2014.895249

Gao, Y., Low, R., Jin, P., & Sweller, J. (2013). Effects of speaker variability on learning foreign-accented English for EFL learners. *Journal of educational psychology*, *105*(3), 649. https://doi.org/10.1037/a0033024

Giannakopoulou, A., Brown, H., Clayards, M., & Wonnacott, E. (2017). High or low? Comparing high and low-variability phonetic training in adult and child second language learners. *PeerJ*, *5*, e3209. https://doi.org/10.7717/peerj.3209

Gonthier, C., Macnamara, B. N., Chow, M., Conway, A. R., & Braver, T. S. (2016). Inducing proactive control shifts in the AX-CPT. *Frontiers in psychology*, *7*, 1822. https://doi.org/10.3389/fpsyg.2016.01822

Halperin, J. M., Sharma, V., Greenblatt, E., & Schwartz, S. T. (1991). Assessment of the Continuous Performance Test: Reliability and validity in a nonreferred sample. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *3*(4), 603. https://doi.org/10.1037/1040-3590.3.4.603

Hardison, D. M. (2003). Acquisition of second-language speech: Effects of visual cues, context, and talker variability. *Applied Psycholinguistics*, *24*(4), 495. https://doi.org/10.1017/S0142716403000250

Heald, S. L., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in systems neuroscience*, *8*, 35. https://doi.org/10.3389/fnsys.2014.00035

Heeren, W. F. L., & Schouten, M. E. H. (2010). Perceptual development of the Finnish /t-t:/ distinction in Dutch 12-year-old children: A training study. *Journal of Phonetics, 38*(4), 594–603. https://doi.org/10.1016/j.wocn.2010.08.005

Huensch, A. (2016). Perceptual phonetic training improves production in larger discourse contexts. *Journal of Second Language Pronunciation*, *2*(2), 183–207. https://doi.org/10.1075/jslp.2.2.03hue

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

Jamieson, D. G., & Morosan, D. E. (1986). Training non-native speech contrasts in adults: Acquisition of the English/ð/-/θ/contrast by francophones. *Perception & psychophysics*, *40*(4), 205-215. https://doi.org/10.3758/BF03211500

Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.

Kartushina, N., & Martin, C.D. (2019), Talker and Acoustic Variability in Learning to Produce Nonnative Sounds: Evidence from Articulatory Training. *Language Learning,* 69: 71-105. https://doi.org/10.1111/lang.12315

Koch, I., Lawo, F., Fels, J. & Vorlaender, M. (2011). Switching in the Cocktail Party: Exploring Intentional Control of Auditory Selective Attention. *Journal of Experimental Psychology: HPP, 37,* 1140 – 1147. https://doi.org/10.1037/a0022189

Lakens, D. (2020, March 29). Effect Sizes and Power for Interactions in ANOVA Designs. Retrieved November 12, 2020, from http://daniellakens.blogspot.com/2020/03/effect-sizes-and-power-for-interactions.html

Lambacher, S., Martens, W., Kakehi, K., Marasinghe, C., & Molholt, G. (2005). The effects of identification training on the identification and production of American English vowels by native speakers of Japanese. *Applied Psycholinguistics, 26*(2), 227-247. https://doi.org/10.1017/S0142716405050150

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

Lee, C. Y., Tao, L., & Bond, Z. S. (2009). Speaker variability and context in the identification of fragmented Mandarin tones by native and non-native listeners. *Journal of Phonetics*, *37*(1), 1-15. https://doi.org/10.1016/j.wocn.2008.08.001

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods, 44,* 325-343. https://dx.doi.org/10.3758%2Fs13428-011-0146-0

Lev-Ari, S. (2018). The influence of social network size on speech perception. *Quarterly Journal of Experimental Psychology, 71*(10), 2249–2260. https://doi.org/10.1177/1747021817739865

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3 Pt 1), 1242–1255. https://doi.org/10.1121/1.408177

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886. https://doi.org/10.1121/1.1894649

Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 351–377). York Press.

MacKain, K. S., Best, C. T., & Strange, W. (1981). Categorical perception of English/r/and/l/by Japanese bilinguals*. Applied Psycholinguistics, 2*(4), 369-390. https://doi.org/10.1017/S0142716400009796

Magnuson, J. S., Yamada, R. A., Tohkura, Y. I., & Bradlow, A. R., Lively, S.E. (1995). Testing the importance of talker variability in non native speech contrast training. *The Journal of the Acoustical Society of America, 97*(5), 3417-3417. https://doi.org/10.1121/1.412450

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 676. https://doi.org/10.1037//0278-7393.15.4.676

Masicampo, E.J., & D. R. Lalande. (2012). A peculiar prevalence of *p* values just below .05. *The Quarterly Journal of Experimental Psychology, 65*(11), 2271-2279. https://doi.org/10.1080/17470218.2012.711335

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, R. H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Milne, A. E., Bianco, R., Poole, K.C., Zhao, S., Oxenham, A.J., Billig, A.J., Chait, M. (2020). *Behavior Research Methods.* https://doi.org/10.3758/s13428-020-01514-0

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. https://doi.org/10.1121/1.397688

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577. https://doi.org/10.1037/xhp0000049

Nishi, K., & Kewley-Port, D. (2007). Training Japanese Listeners to Perceive American English Vowels: Influence of Training Sets. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1496–1509. https://doi.org/10.1044/1092-4388(2007/103)

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472. https://doi.org/10.1121/1.3593366

Potter, C. E., & Saffran, J. R. (2017). Exposure to multiple accents supports infants' understanding of novel accents. *Cognition*, *166*, 67-72. https://doi.org/10.1016/j.cognition.2017.05.031

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Department of Applied Mathematics and Theoretical Physics, Cambridge England, *Technical report NA2009/06*.

Quam, C., Clough, L., Knight, S. and Gerken, L. (2021), Infants' discrimination of consonant contrasts in the presence and absence of talker variability. *Infancy*, *26*: 84-103. https://doi.org/10.1111/infa.12371

R Core Team (2018). R: A language and environment for statistical computing. *R Foundation for Statistical Computing,* Vienna, Austria. https://www.R-project.org/

Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The Effects of Feature-Label-Order and Their Implications for Symbolic Learning. *Cognitive Science, 34*(6), 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Raviv, L., Lupyan G. & Green S. (2022). How variability shapes learning and generalization. *Trends in Cognitive Sciences, 26(*6), 462-483*.*

Riccio, C. A., Reynolds, C. R., Lowe, P., & Moore, J. J. (2002). The continuous performance test: a window on the neural substrates for attention?. *Archives of Clinical Neuropsychology*, *17*(3), 235-272. https://doi.org/10.1016/S0887-6177(01)00111-1

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology*, *70*(5), 897-905. https://doi.org/10.1080/17470218.2015.1136656

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638. https://doi.org/10.1037/0033-2909.86.3.638

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608-635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome Jr, E. D., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology*, *20*(5), 343. https://doi.org/10.1037/h0043220

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic bulletin & review*, *21*(2), 301-308. https://doi.org/10.3758/s13423-014-0595-4

Rouder, J. & Haal, J. M. (2019). Optional Stopping and the Interpretation of The Bayes Factor. *PsyArXiv*. https://doi.org/10.31234/osf.io/m6dhw

Sadakata, M., & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates. *The Journal of the Acoustical Society of America*, *134*(2), 1324–1335. https://doi.org/10.1121/1.4812767

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. *Frontiers in Psychology*, *5*(November), 1–15. https://doi.org/10.3389/fpsyg.2014.01318

Saito, K., Sun, H., and Tierney, A. (2020). Test-Retest Reliability of Explicit Auditory Processing Measures. *bioRxiv.* https://doi.org/10.1101/2020.06.12.149484

Saito, K., Suzukida, Y., Tran, M. and Tierney, A. (2021), Domain General Auditory Processing Partially Explains Second Language Speech Learning in Classroom Settings: A Review and Generalization Study. *Language Learning.* https://doi.org/10.1111/lang.12447

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187-224. https://doi.org/10.1017/S0142716417000418

Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic bulletin & review*, *25*(1), 128-142. https://doi.org/10.3758/s13423-017-1230-y

Shinohara, Y. (2021). Audiovisual English /r/−/l/ Identification Training for Japanese-Speaking Adults and Children. *Journal of Speech, Language, and Hearing Research, 64*(7), 2529-2538. https://doi.org/10.1044/2021_JSLHR-20-00506

Sinkevičiūtė, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, *41*(4), 795-820. https://doi.org/10.1017/S0272263119000263

Sommers, M. S., & Barcroft, J. (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, *119*(4), 2406-2416. https://doi.org/10.1121/1.2171836

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*(2), 231. https://doi.org/10.1017/S0142716407070129

Sommers, M. S., & Barcroft, J. (2011). Indexical information, encoding difficulty, and second language vocabulary learning. *Applied Psycholinguistics*, *32*(2), 417. https://doi.org/10.1017/S0142716410000469

Stacey, P. C., & Summerfield, A. Q. (2007). Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech. *The Journal of the Acoustical Society of America*, *121*(5 Pt1), 2923–2935. https://doi.org/10.1121/1.2713668

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of/rl/by Japanese adults learning English. *Perception & psychophysics*, *36*(2), 131-145. https://doi.org/10.3758/BF03202673

Thomson, R. I. (2018). High variability [pronunciation] training (HVPT): A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation, 4*(2), 208-231. https://doi.org/10.1075/jslp.17038.tho

Wiener, S., Ito, K., & Speer, S. R. (2018). Early L2 Spoken Word Recognition Combines Input-Based and Knowledge-Based Processing. *Language and Speech, 61*(4), 632–656. https://doi.org/10.1177/0023830918761762

Wiener, S., Chan, M.K.M. and Ito, K. (2020), Do Explicit Instruction and High Variability Phonetic Training Improve Nonnative Speakers' Mandarin Tone Productions?. *The Modern Language Journal, 104*: 152-168. https://doi.org/10.1111/modl.12619

Wong, J. W. S. (2012). Training the Perception and Production of English /e/ and /ae/ of Cantonese ESL Learners: A Comparison of Low vs. High Variability Phonetic Training. *14th Australasian International Conference on Speech Science and Technology*, (December), 37–40. Retrieved from http://assta.org/sst/SST-12/SST2012/PDF/INDEXSCR.PDF

Wong, J. W. S. (2014). The Effects of High and Low Variability Phonetic Training on the Perception and Production of English Vowels / e / - / æ / by Cantonese ESL Learners with High and Low L2 Proficiency Levels. *Proceedings of the 15 Th Annual Conference of the International Speech Communication Association*, 524–528. Retrieved from https://repository.hkbu.edu.hk/hkbu_staff_publication/6234

Woods, K. J., & McDermott, J. H. (2015). Attentive tracking of sound sources. *Current Biology*, *25*(17), 2238-2246. http://dx.doi.org/10.1016/j.cub.2015.07.043

Xie, X., Liu, L., & Jaeger, T. F. (2021) Cross-talker generalization in the perception of non-native speech: a large-scale replication. *Journal of Experimental Psychology General, 150*(11), e22–e56. https://doi.org/10.1037/xge0001039

Yamada, R. A. (1993). Effect of extended training on /r/ and /l/ identification by native speakers of Japanese. The Journal of the Acoustical Society of America, 93(4), 2391–2391. https://doi.org/10.1121/1.406052

Zhang, X., Cheng, B., & Zhang, Y. (2021) The Role of Talker Variability in Nonnative Phonetic Learning: A Systematic Review and Meta-Analysis. Journal of Speech, Language, and Hearing Research, 64, 4802-4825. https://doi.org/10.1044/2021_JSLHR-21-00181

Zhang, K., Peng, G., Li, Y., Minett, J.W., & Wang, W.S. (2018). The Effect of Speech Variability on Tonal Language Speakers' Second Language Lexical Tone Learning. *Frontiers in Psychology.* 9. 1982. https://doi.org/10.3389/fpsyg.2018.01982