The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones

Hanyu Dong Corresp., [1] , **Meghan Clayards** [2] , **Helen Brown** [3] , **Elizabeth Wonnacott** Corresp. [1]

[1] Division of Psychology and Language Sciences, University College London, University of London, London, United Kingdom
[2] Department of Linguistics, School of Communications Sciences and Disorders, McGill University, Montreal, QC, Canada
[3] Department of Psychology, Nottingham Trent University, Nottingham, United Kingdom

Corresponding Authors: Hanyu Dong, Elizabeth Wonnacott
Email address: hanyu.dong.10@ucl.ac.uk, e.wonnacott@ucl.ac.uk

High variability training has been found more effective than low variability training in learning various non-native phonetic contrasts. However, little research has considered whether this applies to the learning of tone contrasts. The only two relevant studies suggested that the effect of high variability training depends on the perceptual aptitude of participants (Perrachione, Lee, Ha, & Wong, 2011; Sadakata & McQueen, 2014). The present study extends these findings by examining the interaction between individual aptitude and input variability using natural, meaningful L2 input (both previous studies used pseudowords). Sixty English speakers took part in an eight session phonetic training paradigm. They were assigned to high/low/high-blocking variability training groups and learned real Mandarin tones and words. Individual aptitude was measured following previous work. Learning was measured using one discrimination task, one identification task and two production tasks. All tasks assessed the generalisation of learning. Overall, all groups improved in both production and perception of tones which transferred to novel voices and items, demonstrating the effectiveness of training despite the increased complexity compared with previous research. Although the low variability group exhibited an advantage with the training stimuli, there was no evidence that the different variability training led to different performance in any of the tests of generalisation. Moreover, although aptitude significantly predicted performance in discrimination, identification and training tasks, no interaction between individual aptitude and variability was revealed. We discuss these results in light of previous findings.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones

Hanyu Dong[1], Meghan Clayards[2], Helen Brown[3], & Elizabeth Wonnacott[1]

[1]*Division of Psychology and Language Sciences, University College London, London, UK*

[2]*Department of Linguistics, School of Communications Sciences and Disorders, McGill University, Montreal, QC, Canada*

[3] *Department of Psychology, Nottingham Trent University, Nottingham, UK*

Correspondence concerning this article should be addressed to Elizabeth Wonnacott, Division of Psychology and Language Sciences, Chandler House, 2 Wakefield Street, London, WC1N 1PF.

Email: e.wonnacott@ucl.ac.uk

Declarations of interest: None

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

Abstract

High variability training has been found more effective than low variability training in learning various non-native phonetic contrasts. However, little research has considered whether this applies to the learning of tone contrasts. The only two relevant studies suggested that the effect of high variability training depends on the perceptual aptitude of participants (Perrachione, Lee, Ha, & Wong, 2011; Sadakata & McQueen, 2014). The present study extends these findings by examining the interaction between individual aptitude and input variability using natural, meaningful L2 input (both previous studies used pseudowords). Sixty English speakers took part in an eight session phonetic training paradigm. They were assigned to high/low/high-blocking variability training groups and learned real Mandarin tones and words. Individual aptitude was measured following previous work. Learning was measured using one discrimination task, one identification task and two production tasks. All tasks assessed the generalisation of learning. Overall, all groups improved in both production and perception of tones which transferred to novel voices and items, demonstrating the effectiveness of training despite the increased complexity compared with previous research. Although the low variability group exhibited an advantage with the training stimuli, there was no evidence that the different variability training led to different performance in any of the tests of generalisation. Moreover, although aptitude significantly predicted performance in discrimination, identification and training tasks, no interaction between individual aptitude and variability was revealed. We discuss these results in light of previous findings.

Keywords: Phonetic training; L2 phonetic contrasts; Lexical tone learning

1 Introduction

One challenging aspect of learning a second language (L2) is learning to accurately perceive non-native phonetic categories. This task is particular difficulty where the L2 contains the same acoustic properties as the first language (L1), but used differently (Bygate, Swain, & Skehan, 2013), suggesting that it is challenging to adjust existing acoustic properties in the L1 to learn new L2 categories. This challenge is compounded by the fact that speech is highly variable in the natural linguistic environment. Variability comes not only from the phonetic context but also from differences between speakers. Thus, learners must learn to distinguish the new L2 categories despite all the variability present in the learning input. There is evidence that native listeners can process this variability in speech faster and more accurately than non-native listeners (Bradlow & Pisoni, 1999), indicating that it is indeed a challenge for L2 learners. Despite this, it has been suggested that input variability may be beneficial for second language learning and generalization (Barcroft & Sommers, 2005; Lively, Logan & Pisonni, 1993). However recent evidence suggests that the ability to benefit from variability may depend on individual learner aptitude (Perrachione, Lee, Ha, & Wong, 2011; Sadakata & McQueen, 2014), at least in the learning of lexical tones i.e. the distinctive pitch patterns carried by the syllable of a word which, in certain languages, distinguish meaningful lexical contrasts. The current paper further explores how and when variability supports or impedes learning of new L2 phonetic categories, focusing on English learners of Mandarin tone contrasts.

1.1 High Variability L2 Phonetic Training for Non-Tonal Contrasts

A substantial body of literature has explored whether phonetic training can be used to improve identification and discrimination of non-native phonetic contrasts in L2 learners. An early study by Strange and Dittman (1984) attempted to train Japanese speakers on the English

/*r*/- /*l*/ distinction, a phoneme contrasts that does not exist in Japanese. This training study used a discrimination task in which participants made same–different judgments about stimuli from a synthetic *rock-lock* continuum, receiving immediate trial-by-trial feedback. Participants were given a variety of discrimination and identification tasks pre- and post-training. The key result was that although performance increased both for trained items on the synthesized *rock-lock* continuum, and for novel items on a synthesized *rake-lake* continuum, participants failed to show any improvement for naturally produced minimal pair speech tokens. Later research suggested that a key factor which prevented generalization to natural speech tokens was a lack of variability in the training materials: Variability was present in the form of the ambiguous intermediate stimuli along the continuum, however, there was a single phonetic context and a single (synthesized) speaker. Logan, Lively, and Pisoni (1991) also trained Japanese learners on the English /*r*/-/*l*/ contrast, but included multiple natural exemplars (67 minimal pairs, where the target speech sounds appeared in different phonetic contexts) and multiple speakers (four males and two females). Their pre- and post- training tests involved novel and trained words spoken by both trained and novel speakers. In contrast to Strange and Dittman, they found that participants successfully generalized to both new speakers and new words. This was the first study to indicate the importance of variability within the training material. A follow up study by Lively, Logan, and Pisoni (1993) provided further evidence for this by contrasting a condition with *high variability* input with one with *low variability* input in which the stimuli were spoken by a single speaker (although still exemplified in multiple phonetic environments). Participants in the low variability group improved during the training sessions but failed to generalise this learning to new speakers.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

Following Logan et al. (1993) the use of high variability training materials has become standard in L2 phonetic training – the so called *"high variability phonetic training"* (HVPT) methodology. This methodology has been successfully extended to training a variety of contrasts in various languages such as learning of the English /*u:*/-/ʊ/ distinction by Catalan/Spanish bilinguals (Aliaga-García & Mora, 2009), learning of the English /*i:*/-/ɪ/ contrasts by native Greek speakers (Lengeris & Hazan, 2010; Giannakopoulou, Uther & Ylinen, 2013), and learning of the English /*w*/-/*v*/ distinction by native German speakers (Iverson, Ekanayake, Hamann, Sennema, & Evans, 2008).

There is also some evidence that this type of perceptual training benefits production in addition to perception. Bradlow, Pisoni, Akahane-Yamada, and Tohkura (1997) found that production of the /*r*/-/*l*/ contrast improved in Japanese speakers following HVPT, with this improvement being retained even after three months. Similar improvement on the production of American English mid to low vowels by Japanese's speakers following HVPT was also reported by Lambacher, Martens, Kakehi, Marasinghe, and Molholt (2005). However, the evidence here is mixed: a recent study (Alshangiti & Evans, 2014) employed HVPT to train Arabic learners on non-native English vowel contrasts and found no improvements in production, although participants receiving additional explicit production training did show some limited improvement.

The finding that variability boosts generalization is intuitively sensible: Experience of variation allows the formation of generalized representations that include only phonetically relevant cues and exclude irrelevant speaker identity cues. However it is notable that the seminal experiments of Logan and colleagues had a small sample (the tests of generalization were administered to only three of the participants in Logan et al. 1991), and since this work, relatively few studies have explicitly tested the benefit of high variability training by directly comparing high variability and low variability training conditions. Clopper and Pisoni (2004) found a benefit of high variability, although this focused on dialect categorization rather than L2 phonetic learning. They tested participants' ability to categorize dialects following exposure to high variability training (three speakers per dialect) compared with low variability training (one speaker per dialect), finding better generalization after high variability training. Sadakata and McQueen (2013) trained native Dutch speakers with geminate and singleton variants of the Japanese fricative /*s*/. Participants were trained with either a limited set of words recorded by a single speaker (low-variability) or with a more variable set of words recorded by multiple speakers (high-variability). Critically, the total amount of exposure to the contrast was held constant across conditions such that each item in the low-variability condition was repeated more frequently than each item in the high-variability condition. Both types of training led to increases in both the identification and discrimination of the novel contrast, including generalization to untrained fricatives and speakers, however for the identification task the improvement was greater following high variability training.

More recently, Giannakopoulou, Brown, Clayards, and Wonnacott (2017) compared matched high variability (four speakers) and low variability (one speaker) training for adult and child (8 year old) native Greek speakers who were trained on the English /*i:*/-/*i*/ contrast. In contrast to the results of Logan et al. (1993), this study did *not* show a benefit for high variability compared to low variability training in either age group, even for generalization items. However, for adult participants, it is unclear the extent to which this was due to ceiling effects. Two other previous studies which specifically manipulated variability during learning of novel phonetic categories are those by Perrachione, Lee, Ha, & Wong (2011) and Sadakata and McQueen

(2014) which both looked at the learning of lexical tone. We discuss these studies in more detail in the following section.

Finally, there is also evidence of a benefit of high variability training in L2 vocabulary learning: With more varied training materials, (either multiple speakers or multiple voice quality types) participants show greater learning in both production and reception tests (Barcroft & Sommers, 2005, 2014; Sommers & Barcroft, 2007, 2011).

1.2 Phonetic Training of L2 Lexical Tones

Each of the phonetic training studies discussed above involved training a *segmental* contrast (consonantal or vocalic). Another type of phonological contrast which exists in some natural languages is lexical tone, whereby the pitch contour is used to distinguish lexical or grammatical meanings (Yip, 2002). For example, Mandarin Chinese has four lexical tones; level-tone (Tone 1), rising-tone (Tone 2), dipping tone (Tone 3) and falling-tone (Tone 4). These pitch contours combine with syllables to distinguish meanings. For instance, the syllable *ba* combines with the four tones to mean: eight (*bā*, Tone 1), pluck (*bá*, Tone 2), grasp (*bǎ*, Tone 3) and father (*bà*, Tone 4). Each of these words thus forms a minimal pair with each of the others. Note that while languages such as English use pitch information extensively for intonation – such as forming a question or for emphasis – they do not use pitch information lexically, causing difficulties for learners of Mandarin as an L2.

The first study examining lexical tone training was conducted by Wang, Spence, Jongman, and Sereno (1999). A similar paradigm to that used by Logan et al. (1991) was adopted using four speakers for training. Training consisted of a two-alternative forced choice (2AFC) task in which participants heard a syllable whilst viewing two standard diacritic representations (i.e., →, ↗, ∨, ↘, which are iconic in nature). They were asked to pick out the picture of the arrow that corresponded to the tone and received feedback. At test, participants chose which tone they had heard out of a choice of all four (4AFC task). There were also two generalisation tasks, one with 60 new words produced by one of the training speakers, and the other with an additional 60 new words produced by a new speaker. Training materials were all real monosyllabic Mandarin words that varied in the consonants, vowels and syllable structure. Native speakers of American English showed significant improvement in the accuracy of tone identification after eight sessions of high variability training over two weeks and this generalized to both new words and new speakers.

In a follow up study, Wang, Jongman and Sereno (2003) used the same training paradigm to test whether learning transferred to production. They recruited participants taking Mandarin courses and asked them to read through a list of 80 Mandarin words written in Pinyin (an alphabetic transcription) before and after training. These production were rated by 82 native Mandarin speakers blind to whether each recording was from pre- or post-test. They found improvements in production, although these were mainly seen in pitch height rather than pitch contour.

These studies suggested that as with segmental phoneme contrasts, high variability training could also facilitate the learning of tone contrasts. However, Wang and colleagues (1999, 2003) used only HVPT. Following the results of Logan et al. (1991, 1993) there is an interest in exploring whether high variability training has an advantage over low variability training. The first study to investigate this for the training of lexical tone was conducted by Perrachione et al. (2011). They trained native American English speakers with no previous knowledge of Mandarin (or any other tonal language), using English monosyllabic pseudowords combined with Mandarin tones 1 2, and 4. The training task used either low variability (one speaker) or high variability (four speaker) input. The pseudowords were associated with concrete objects displayed in pictures. During the training, participants matched the sound they heard with one of three pictures presented, where the three words associated with these pictures were minimal trios that differed only in tone. They received feedback on a trial-by-trial basis. Learning was tested using a version of the training task with new talkers (and with feedback removed). Importantly, Perrachione et al. (2011) were also interested in the role of individual differences in learning. Therefore, in addition to the key tests of the training materials, they also determined participants' baseline ability to perceive the tone contrasts using a *Pitch Contour Perception Test (PCPT).* In this task, participants heard a vowel produced with either Mandarin tone 1, 2 or 4 whilst viewing pictures of the three standard diacritics, and were asked to pick out the picture of the arrow that corresponded to the tone. Based on performance in this task before training, the researchers grouped participants into high and low aptitude groups. The key finding of this study was that while the low variability group outperformed the high variability group during training (presumably due to accommodation to a repeated speaker through the task), there were no differences between the high and low variability groups during test, even though test items involved novel speakers and thus probed generalization. Critically however, there was an interaction between individuals' aptitude categorization (as defined by the PCPT) and the type of variability training: Only the participants with high aptitude benefited from high variability training, while those with low aptitude actually benefited more from low variability training.

Another training study by Sadakata and McQueen (2014) also explored the relationship between input variability and individual aptitude in lexical tone training, though using rather different training and testing materials. They trained native Dutch speakers (with no prior knowledge of Mandarin or any other tonal language) using naturally produced bisyllabic

Mandarin pseudowords. The two syllables in each word either had Tone 2 followed by Tone 1, or Tone 3 followed by Tone 1, and each tone pair was randomly assigned one of two numeric labels (1, 2 - so for example for one participant Tone 2-Tone 1 was labelled "1", Tone 3-Tone 1 was labelled "2"). During the training task, participants were asked to identify the tone pair type of each stimulus by choosing the correct numeric label (e.g. hear /pasa/ with Tone 2-Tone 1, correct response is 2). Thus, in contrast to the study by Perrachione et al. (2011), participants did not need to learn the meaning of each word. Input variability was manipulated, with three levels (low/medium/high). In contrast to the work by Perrachione et al., where the high variability and low variability conditions differed only in terms of the number of speakers, in this study variability was increased both by including more speakers and more items (pseudowords). The test session used a similar design to the training sessions but included a 3AFC test (to prevent ceiling effect, a new untrained tone pair [Tone 1 – Tone 1], was included alongside the trained contrasts and assigned a new numeric label ("3")).

As in the study by Perrachione et al. (2011), Sadakata and McQueen (2014) also tested individual aptitude but with a different method. They employed a categorization task using stimuli from a six step Tone 2 to Tone 3 continua (created using natural productions of the two tones with the Mandarin vowel /*a*/ as endpoints and linearly interpolating between these endpoints). Participants were asked to identify if the sound they heard was more like Tone 2 or Tone 3 and a categorization slope was obtained for each participant, providing a measure of their ability to discriminate this contrast (which is generally found to be the most challenging tone contrast for L2 learners of Mandarin). Participants were grouped according to their slopes, and as in Perrachione et al., this grouping was entered as a factor in the analyses of the main test of learning. The results were similar to those of Perrachione et al.: there was no group level benefit of high variability training but instead an interaction between individual aptitude and variability condition, which was due to the fact that only participants with high aptitude benefited from high variability training, while those with lower aptitude actually benefitted more from low variability training. There was also no interaction between aptitude and variability condition in the tests of generalization to new speakers or items.

The results of these studies thus provide mutually corroborating evidence – using somewhat different training and testing methods - that the ability to learn from high variability input is dependent on learner aptitude. Perrachione et al. (2011) suggest that one reason why low aptitude participants may struggle with multi-speaker input is that the speakers were intermixed during training: This requires trial-by-trial adaption to each speaker, which was not required in the corresponding single speaker low variability conditions. This may place a burden on learners (see Nusbaum & Morin, 1992; Mattys & Wiget, 2011 for evidence that intermixed multi-speaker stimuli are difficult even for L1 processing and that this interacts with constraints on working memory and attention). To test this, Perrachione et al. included a second experiment in which items from each speaker were presented in separate blocks (as is more common in high variability phonetic training). This improved performance with trained items compared with unblocked training for low aptitude learners only, confirming the hypothesis that switching between speakers interferes with learning for low aptitude learners. On the other hand, Sadaka and McQueen (2014) employed a blocked presentation in their high variability condition, so that trial-by-trial inconsistency *cannot* explain the greater difficulty of low aptitude learners in this study.

1.3 The Current Study

The finding that learning from multiple voices is more or less effective for different groups of learners may have implications for those interested in designing training tools for educational purposes. The fact that the effect has been found using quite different methods is encouraging. Here we further probe this finding in a new paradigm in which naive participants are trained using natural, meaningful stimuli from Mandarin Chinese. The current study serves as a partial replication and extension of the two previous studies by Perrachione et al. (2011) and Sadakata and McQueen (2014).

There are three important points to note with regards to our methodology. First, we trained participants on real Mandarin words produced by native speakers. This stands in contrast to previous studies which have trained participants only on pseudowords: Perrachione et al. (2011) used Mandarin tones with English pseudowords, whilst Sadakata and McQueen (2014) used Mandarin pseudowords. Second, while previous studies have trained participants on only three of the four tones, we trained participants on all four Mandarin tones (six tone contrasts) given that learners of Mandarin will need to learn the complete set. Thirdly, we embedded tone learning in a vocabulary learning task. This contrasts with the procedure used by Sadakata and McQueen, where participants were trained to map tonal categories onto (arbitrary) numbers, as well as with other HVPT studies in which participants were trained to map phonetic categories to orthographic categories (e.g. "*r*" and "*l*", Logan et al. 1993). However the procedure is in line with that used by Perrachione et al. (described above), where participants were trained to associate pseudowords containing tonal information with pictures of common objects such as table, bus, or phone. Learning both tones and lexical items simultaneously more closely resembles real world L2 learning situations.

The key manipulation in the current study was the amount and type of variability that occurred during training. Following Perrachione et al. (2011), we compared training given to different groups of learners: low variability training (one speaker), high variability training (four speakers intermixed within each training session) and high variability blocking training (four speakers each presented in separate blocks). We predicted that the difficulty of high variability input for lower aptitude participants would be greater in the unblocked condition, thus potentially increasing the likelihood of seeing the predicted interaction between variability and learner aptitude. On the other hand, blocked input is more usual of HVPT (e.g. Logan et al. 1991; Iverson, Hazan & Bannister, 2005) and may increase the possibility of seeing any benefits of speaker variability on generalization.

We used two perceptual tasks designed to tap individual aptitude. These were adapted from those used Perrachione et al. (2011) and Sadakata and McQueen (2014). However, while the previous studies grouped participants into one of two categories (high aptitude *vs.* low aptitude) based on the aptitude score, in current study they were used as continuous measures (allowing us to avoid assigning an arbitrary "cut off" for high *vs*. low aptitude groups, and the loss of information which occurs when an underlying continuous variable is turned into a binary measure). Note that the statistical approach used in this paper (logistic and linear mixed effect models) allowed us to include continuous predictors and look at their interactions with other factors.

We also included several measures of learning. The three interval oddity task required participants to pick out the "different word" after hearing three words spoken aloud. The three words were minimal triplets but with only two tone used (e.g. *bā,* Tone 1; *bā,* Tone 1; *bà*, Tone 4). Both speaker novelty and item novelty were manipulated. The word repetition task, in which participants repeated spoken Mandarin words, provided a test of production which could be conducted both pre and post-test. Item novelty was again manipulated. In the post-test session only, we included two additional tests: a picture identification test and a picture naming task. The picture identification test was similar in form to the training session (2AFC picture identification), however new speakers were used in order to test speaker generalization. The picture naming task required participants to name the pictures used in training in Mandarin. Note that last two tasks test both the ability to perceive/produce the tone distinctions in Mandarin, but also to link these to meaning, potentially tapping more directly in to mechanisms relevant to word learning.

In sum, the following experiment assessed whether individuals' aptitude would interact with high/low variability training. It used real Mandarin stimuli with all four Mandarin tones embedded in a vocabulary learning task, and included tests of both perception and production.

2 Method

2.1 Participants

Sixty adults recruited from UCL Psychology Subject Pool participated in the experiment, twenty in each of the three conditions (low variability, high variability, high variability blocking). Participant information is summarised in *Table 1*. There was no difference between these groups in age, F (2, 57) = 1.95, p = .15. Participants had no known hearing, speech, or language impairments. Written consent was obtained from participants prior to the first session. Each participant was paid £45 at the end of the study.

All participants except three were native speakers of English. Of these three, one participant (low variability condition) was a native bilingual of English and Hindi, one participant (high variability condition) was a native French speaker, and one participant (high variability condition) was a native Finnish speaker. Critically none had any prior experience of Mandarin Chinese or any other tonal language. On average, participants learned 2.4 (*SD* = 0.8) languages and the average age for starting to learn the first L2 was 12.6 (*SD* = 1.3).

2.2 Stimuli

2.2.1 Stimuli used in Training and in the Picture Identification, Three Interval Oddity, Word Repetition and Picture Naming Tests

These stimuli consisted of 36 minimal pairs of Mandarin words (6 minimal pairs for each of the six tone contrasts for each of the four Mandarin tones). The words in each pair contained the same phonemes, differing only in tones (e.g. *māo*, Tone 1 [*cat*] vs. *mào*, Tone 4 [*hat*]). The words were chosen to be picturable and to start with a wide range of phonemes (see Appendix A). In order to examine generalization across items, half of the word pairs (3 per tone contrast) were designated "trained” words and used in both training and testing: the other half were designated "untrained" words and were encountered only at test.

The full set of 72 Mandarin words was recorded by two groups of native Mandarin speakers using a Sony PCM-M10 handheld digital audio recorder. The first group was made up of three female speakers and two male speakers, (F1, F2, F3, M1, M2). These stimuli were used in the training, word repetition and picture identification tasks. The second group consisted of three new female speakers and two new male speakers (FN1, FN2, FN3, MN1, MN2). These stimuli were used in the Three interval oddity task (making all new speakers in that task). Table 2 summarises how speakers were assigned to each task.

In the low variability condition only one speaker (Trained voice 1) was used in training, and this same speaker was also used as the test voice in the Word Repetition test and for trained test items in the Picture Identification test. In the high variability condition, four speakers were used in training. Only one of these speakers (Trained voice 1) was used in the Word Repetition test and for trained items in the Picture Identification test (the same speaker across both tests). In both conditions, a further speaker (New voice 1) was assigned to the untrained test items in the Picture Identification test. The assignment of speakers was rotated across participants, resulting in 5 counterbalanced versions of each condition (see Table 2). This ensured that any difference found between the low and high variability conditions, and between trained and new voices, were not due to idiosyncratic difference between voices. There was no counterbalancing of speaker in other tasks.

All words were edited into separate sound files, and peak amplitude was normalised using Audacity (Audacity team, 2015, http://audacity.sourceforge.net/). Any background noise was also removed. All recordings were perceptually natural and highly distinguishable as judged by native Chinese speakers. Clipart pictures of the 72 words were selected from free online clipart databases.

2.2.2 Stimuli used in the Aptitude Tests:

Pitch Contour Perception Test: Six Mandarin vowels (/*a*/, /*o*/, /*e*/, /*i*/, /*u*/, /*y*/) were repeated in the four Mandarin tones by two male and two female native Mandarin speakers (MN1, MN2, FN1, FN2 from taker set 2) making 96 stimuli in total. Stimuli were identical across conditions and participants.

Categorization of Synthesized Tonal Continua: Natural endpoints were chosen from a native Mandarin male speaker producing the word '*wan*' with both Tone 2 and Tone 3. A neutral vowel was also recorded by a native male English speaker producing the 'father vowel' /a/. This vowel was edited slightly to remove portions containing creaky voice at the end.

The three syllables (*wan* [Tone 2], *wan* [Tone 3], /*a*/) were then manipulated in Praat (Boersma & Weenink, 2015). All three syllables were normalized to be approximately 260 ms long using the PSOLA method. The neutral vowel was manipulated to have a flat pitch (148 Hz) and a flat intensity contour (75dB). The pitch contours of the two natural endpoints were extracted and a 6-step pitch continuum (Step 1: Tone 2, Step 6: Tone3) was generated by linearly interpolating between the endpoints. These six pitch contours were then each superimposed on a copy of the neutral vowel using the PSOLA method. Stimuli were identical across participants and conditions.

2.3 Procedure

The experiment involved three stages (see *Figure 2.3*): Pre-test (session 1), training (sessions 2-7), and post-test (session 8). Participants were required to complete all eight sessions within two weeks, with the constraint of one session per day at most. The majority of sessions took place in a quiet, soundproof testing room in Chandler House, UCL. The remaining sessions took place in a quiet room in a student house.

Participants were given a brief introduction about the aim of the study and told that they were going to learn some Mandarin tones and words. They were explicitly told that Mandarin has four tones (flat, rising, dipping and falling) and that the tonal differences were used to distinguish meanings. The experiment ran on a on a Dell Alienware 14R laptop with a 14-inch screen. The experiment software was built using a custom-built software package developed at the University of Rochester.

The specific instructions for each task were displayed on- screen before the task started. After each task, participants had the opportunity to take a 1-minute break. The tasks completed in each session are listed in *Figure 2.3* and described in more detail below. Note that the PCPT and CSTC were carried out at the beginning of the first session as they provided the measure of individual aptitude prior to exposure to any Mandarin stimuli. There was no time limit for making responses in any of the tasks. Participants wore a pair of HD 201 Sennheiser headphones throughout the experiment.

2.3.1 The Pitch Contour Perception Test

This test was based on the work of Wong and Perrachione (2007). Participants heard a tone (e.g. /*a*/ [Tone 1]), while viewing pictures of four arrows indicating the different pitch contours on the screen. Participants clicked on the arrow that they thought matched the tone heard. No feedback was provided. There were 96 stimuli in total (4 speakers * 4 tones * 6 vowels). Participants completed this task twice, at both pre- and post-test. The main purpose of this task was to provide a measure of individual differences in tone perception prior to training, following Perrachione et al. (2011). Although Perrachione et al. only conducted this task at pre-test, for consistency with the CSTC (described below) we also repeated the test at post-test and conducted analyses to identify whether performance on this task was itself improved as a result of training (see Section 3.3.2).

2.3.2 Categorization of Synthesized Tonal Continua

This test was based on Sadakata and McQueen (2014). Participants first practiced listening to Tone 2 and Tone 3. They heard the tone while viewing the corresponding picture of an arrow. Each tone was repeated 10 times. Then, for each test trial, participants were asked to decide if the sound they heard was closer to Tone 2 or Tone 3 by clicking on the corresponding arrow. No feedback was provided. The speech continua consisted of 6 steps (Step 1: Tone 2, Step 6: Tone 3). Each of the six steps was repeated 10 times per block. Participants completed two blocks, with an optional 1 minute break in the middle, resulting in 120 trials in total. The main purpose of this task was to provide a measure of individual differences in tone perception prior to training, following Sadakata and McQueen (2014). In line with their procedure, participants completed the task both before and after training and we conducted analyses to explore whether there was improvement from pre to post-test (see Section 3.2.1).

2.3.3 Three Interval Oddity Test

This task required subjects to identify the "different" stimulus from a choice of three Mandarin words. Each of the three words within a trial was spoken by a different speaker. Four speakers were used (3 female, 1 male). All speakers were untrained (i.e., not used during training; see *Table 2*). Each trial used one of the 36 minimal pairs from the main stimuli set (18 trained pairs, 18 untrained pairs). Preliminary work suggested that trials differed in difficulty depending on whether the "different" stimulus was spoken by the single male speaker, or one of the three female speakers. We therefore ensured that there were equal numbers of the following trial types: (i) "Neutral" - all three words were spoken by female speakers (ii) "Easy" - the "different" word was spoken by a male speaker and the other two were spoken by female speakers; (iii) "Hard" - the "different" word was spoken by a female speaker and the other two were spoken by one male speaker and one female speaker. Each of the words in the minimal pair was used once as the target ("different") word, making 72 trials in total.

During the task, three frogs were displayed on the screen. Participants heard three words (played with ISIs of 200ms) and indicated which word was the odd one out by clicking on the appropriate frog, which could be in any of the three positions. They could not make their response until after all three words had been heard, at which point a red box containing the instruction "click on the frog that said the different word" appeared at the bottom of the screen. No feedback was given after each trial. Participants completed this task twice – once in the pre-test, and once in the post-test (see *Figure 2.3*).

2.3.4 Word Repetition Test

All seventy-two Mandarin words from the main stimuli set were presented one at a time in a randomised order. They were always spoken by the same speaker and this speaker was also used in their training stimuli (Training voice 1; see *Table 2*). After each word, two seconds of white noise was played. Participants were instructed to listen carefully to the word and then to repeat the word aloud after the white noise. The white noise was included to make sure that participants had to encode the stimulus they were repeating, rather than relying on the phonological loop, which would be pure imitation (Flege, Takagi & Mann, 1995). Verbal responses were digitally recorded and were later transcribed and rated by native speakers of Mandarin (see Section 3.3.1.1). This task was completed once in the pre-test and once in the post-test.

2.3.5 English Introduction Task

This task was included in case the meaning of some pictures were ambiguous (not all items were concrete nouns – e.g. "*to paint*"). Participants saw each of the 36 pictures from the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

training set presented once each in random order and heard the corresponding English word. No response was recorded. Participants completed this task only once, at the end of the pre-test session.

2.3.6 Training Task

Participants completed the training task in Session 2-7. On each trial, participants heard a Mandarin word and selected one of two candidate pictures displayed on the computer screen. The two picture always belonged to the same minimal pair (see *Figure 2.3.6*). After selecting a picture, the participant was informed whether their answer was correct (a green happy face appeared) or incorrect (a red sad face appeared). If the correct choice was made, a picture of a coin also appeared in a box on the left-hand side of the screen, with the aim of motivating participants to try to earn more coins in each subsequent session of training. After that, everything but the correct picture was removed from the screen and the participant heard the correct word again. In the lower right corner of the screen a trial indicator of X/288 was displayed where X indicated the number of trials completed. This tool helped participants to keep track of their performance (see *Figure 2.3.6*).

There were 18 picture/word pairs used. Each word was used as the target word four times. Thus, each picture pair appeared eight times, resulting in288 trials in total per session. Participants were assigned to one of the following condition: low variability, high variability and high variability blocking (with the assignment of speakers counterbalanced – see *Table 2*). Each session lasted for approximately 30 minutes.

In the low variability condition, only *one* speaker was used. In the high variability condition, *four* speakers were used. For these two condition, all 288 trials were randomized so there was no fixed order of speaker. For each participant, each of their six training sessions was identical. In the high variability blocking condition, the stimuli were the same as those in the high variability condition. However, from Day 1 to Day 4 of training (i.e., Session 2-5), only one speaker was involved on each day's training session, with the trained speaker that was used in the test tasks (e.g. F1 for Version 1) always occurring on Day 3 (i.e., Session 4). On Days 5 and 6 of training (i.e., Sessions 6 and 7), participants heard all four speakers, each in a separate block, each word was repeated twice in each voice on these days. The trained speaker used in the test tasks always occurred in the third block. After each block, the number of coins they had earned so far was displayed on the screen. For each participant, the structure of the training task was identical on Days 5 and 6.

2.3.7 Picture Identification Test

This task was the same as the training task with the following changes. Firstly, each word was only repeated twice, once by a trained speaker (Trained voice 1) and once by an untrained speaker (New voice 1), making 72 trials in total. Secondly, no feedback was given. This task was completed only in the post-test.

2.3.8 Picture Naming Test

All 36 pictures from the training words were presented in a randomised order. Participants were instructed to try to name the picture using the appropriate Mandarin word. Verbal responses were recorded and were later transcribed and rated by native Mandarin speakers (see Section 3.5.2). This task was completed only in the post-test.

2.3.9 Questionnaires

Participants completed a language background questionnaire after the experiment. Participants were asked to list all the places they had lived for more than 3 months and any languages that they had learned. For each language the participant was asked to state: (a) how long they learned the language for and their starting age; (b) to rate their own current proficiency of the language.

3 Results and Discussion

3.1 Statistical Approach

Three different sets of analyses are reported. First, we conducted the analysis on two individual measures: CSTC (Section 3.2.1) and PCPT (Section 3.2.3). The primary aim of these analyses was to ensure that the three groups did not differ at pre-test, however we also looked for possible differences at post-test. Second, separate analyses are reported on: data from the tests administered pre- and post- training (i.e. word repetition task (Section 3.3.1) and Three Interval oddity task (Section 3.3.2), the data collected during training (Section 3.4) and the data from the two tasks administered only at post-test (i.e. the picture identification task (Section 3.5.1) and picture naming task (Section 3.5.2). These analyses, explore the effects of our experimentally manipulated conditions on the various measures of Mandarin tone learning. Third, analyses were conducted exploring the role of aptitude in each of these tasks (Section 3.6). Specifically, we wanted to see whether aptitude interacted with *variability-condition* in predicting the benefits of training, in line with the predictions of previous research (Perrachione et al., 2011; Sadakata & McQueen, 2014).

Except where stated, analyses used logistic mixed effect models (LMEs; Baayen, Davidson, & Bates, 2008; Jaeger, 2008; Quené & van den Bergh, 2008) using the package lme4 (Bates, Maechler, & Bolker, 2013) for the R computing environment (R Development Core Team, 2010). LMEs allow binary data to be analysed with logistic models rather than as proportions, as recommended by Jaeger (2008). In each of the analyses, the factor *variability-condition* has three levels (low variability [LV], high variability [HV], and high variability blocking [HVB]) which we coded into two contrasts with LV as the baseline (LV versus HV, LV versus HVB). An exception to this is the training data, where a model containing all three conditions would not converge and we took a different approach, as described in Section 3.4. We also included the interactions between these contrasts and the other factors. We used centred coding which ensued that other effects were evaluated as averaged over all three levels of *variability-condition* (rather than the reference level of LV[1]). Similarly, in the Three Interval Oddity, we included a *trial-type* factor (to control for the fact that participants were likely to find some trial types easier than others) – this had three levels ((i)"Neutral" - all three words were spoken by female speakers (ii) "Easy" - the "different" word was spoken by the one male speaker (iii) "Hard" - the "different" word was spoken by one of the two female speakers) and for this we included contrasts with neutral ("neutral versus easy" and "neutral versus hard") again using centered coding. In order to perform the analysis comparing pre- and post-test performance, *test-session* was coded as a factor with two levels (pre-test/post-test) with "pre-test" set as the reference level. This allowed us to look at the (accidental) possible differences between the experimental conditions at the pre-test stage, as well as whether post-test performance differed from this baseline. All other predictors, including both discrete factor

[1] This differs from the default coding of contrasts in the lme4 package. It was achieved by replacing the three-way factor "condition" with two centred dummy variables and using the main fixed effects from the output of this model.

codings with two levels (*item-novelty* in the Word Repetition and Three Interval Oddity tasks, and *voice-novelty* in the Picture Identification task) and numeric predictors (*training-session*) in the Training data analyses and the individual difference measures in the models reported in Section 3.7), were centred to reduce the effects of collinearity between main effects and interactions, and in order that main effects were evaluated as the average effects over all levels of the other predictors (rather than at a specified reference level for each factor). We automatically put experimentally manipulated variables and all of their interactions into the model, without using model selection (except for "*trial-type*" in the Three Interval Oddity task which works as a control factor and for this factor we only used its main effect and the interaction with *test-session*). However, we did not inspect the models for all main effects and interactions. Instead, we report statistics which were necessary to look for accidental differences at pre-test, and those related to our hypotheses. We aimed to examine whether the training improves participants' performance on both new items and new voices and whether such improvement was modulated by their individual aptitudes. Participant is included as a random effect and a full random slope structure was used (i.e., by-subject slopes for all experimentally manipulated within-subject effects (*test-session*, *voice-novelty*, *item-novelty*) and interactions, as recommended by Barr, Levy, Scheepers, and Tily, 2013. In some cases the models did not converge and in those cases correlations between random slopes were removed. Models converged with Bound Optimization by Quadratic Approximation (BOBYQA optimization; Powell, 2009). R scripts showing full model details can be found here: https://osf.io/wdh8a/?view_only=d1557462138447ffbaafaf7a59662df8.

3.2 Individual Aptitude Tasks

3.2.1 Categorisation of Synthesized Tonal Continua

We estimated individual's performance on the CSTC task following Sadakata and McQueen (2014). We used the Logistic Curve Fit function in SPSS to calculate a slope coefficient for each participant (Joanisse, Manis, Keating & Seidenberg, 2000). The slope (standardized β) indicates individual differences in tone perception. The smaller the slope, the better the performance. According to Sadakata and McQueen, the data of participants with a slope measure greater than 1.2 were removed from the analysis. Using this threshold 43 out of 60 participants failed the threshold. This is consistent with the observation that most of the participants were not able to consistently categorize the endpoints of the continua, indicating that this was not a good test of aptitude. We do not report further analyses with this aptitude variable however they can be found in the supplemental materials (https://osf.io/wdh8a/?view_only=d1557462138447ffbaafaf7a59662df8).

3.2.2 The Pitch Contour Perception Test

The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were the contrasts between conditions (LV versus HV; LV versus HVB) and *test-session* (pre-test, post-test). (Note - average accuracy in each condition is also included in the table of participant details; Table 1, section 2.1). There was no significant difference between the LV and HV groups at pre-test (β = -0.35, *SE* = 0.26, *z* = -1.38, *p* = 0.17) or between the LV and HVB groups (β = 0.17, *SE* = 0.26, *z* = 0.66, *p* = 0.51) on this measure. Participants showed significant improvement after training (β = 0.21, *SE* = 0.05, *z* = 4.13, $p < 0.001$).

In sum, for this measure of perceptual ability our three participant groups did not differ in their performance and the groups showed equivalent improvement from pre- to post- test. Given that this measure is affected by training, we used participants scores at pre-test as our measure of individual differences in the analyses reported in Section 3.6.

3.3 Tests Administered Pre- and Post- Training

3.3.1 Word Repetition

3.3.1.1 Coding and inter-rater reliability analyses

The same methods were used for both production tests – i.e. the Word Repetition test (pre- and post-) and the Picture naming task (post-test only). The files were combined into a single set, along with the 360 stimuli which were used in the experiment (and which were produced by native Mandarin speakers). The latter items were included in order to examine whether the raters were reliable. All stimuli were rated by two raters: Rater 1 was the first author and Rater 2 was recruited from the UCL MA Linguistics program and was naïve to the purposes of the experiment. Raters were presented with recordings in blocks in a random sequence (blind to test-type, condition, whether the stimulus was from pre-test or post-test and whether it was produced by a participant or was one of the experimental stimuli). For each item, raters were asked to (i) identify the tone, (ii) give a rating quantifying how native-like they thought the pronunciation was compared (1-7 with 1 as not recognizable and 7 as native speaker level), and (iii) transcribe the pinyin (segmental pronunciation) produced by the participants.

Three measurements were taken from the production tasks: mean accuracy of tone identification (Tone accuracy), mean tone rating (Tone rating) and mean accuracy of production of the pinyin (derived by coding each production as correct (1= the entire string is correct) or incorrect (0 = at least one error in the pinyin)). As a first test of rater reliability, performance with the native speaker stimuli was examined– these were near ceiling: Rater 1: Tone accuracy =

98%, Tone rating = 6.7, Pinyin accuracy = 80%; Rater 2: Tone accuracy = 87%, Tone rating = 6.5, Pinyin accuracy = 80%).

Furthermore, for the remaining data (i.e. the experimental data) inter-rater reliability was examined for both measures for the two production tasks. For the binary measures (Tone accuracy and Pinyin accuracy), kappa statistics were calculated using the "fmsb" package in R (Cohen, 2014). For the word repetition data, for Tone accuracy *kappa* = 0.43 ("moderate agreement"), and for Pinyin accuracy *kappa*= 0.33 ("fair agreement"; Landis & Koch, 1977). For the Picture Identification test, for Tone accuracy *kappa* = 0.68 ("substantial agreement") and for Pinyin accuracy *kappa* = 0.54 ("moderate agreement"); For the Tone rating, the package "irr" in R was used to access the intra-class correlation (McGraw & Wong, 1996) based on an average-measures, consistency, two-way mixed-effects model. For Word Repetition, *ICC* = 0.28 and for Picture Identification *ICC* = 0.44; according to Cicchetti (1994), values less than .40 are regarded as "poor". Given this, we do not include analyses with Tone Rating as the dependent variable (though these data are included in the data set https://osf.io/wdh8a/?view_only=d1557462138447ffbaafaf7a59662df8). All of the analyses presented in Sections 3.3.1 and 3.5.2 were based on Rater 2 (the naive rater).

3.3.1.2 Tone accuracy

The predicted variable was whether a correct response was given (1/0) on each trial (as identified by the coder). The predictors were *test-session* (pre-test, post-test), *variability-condition* (LV versus HV, LV versus HVB) and *item-novelty* (trained, untrained). The mean accuracy, split by test session and training condition, is shown in *Figure 3.3.1.2*.

At pre-test, there was no significant difference between the LV and the HV group (*β* = -0.09, *SE* = 0.20, *z* = -0.46, *p* = .65) nor between the LV and the HVB group (*β* = 0.05, *SE* = 0.20, *z* = 0.27, *p* = .79), suggesting the groups started at a similar level. There was also no difference between trained and untrained words at pre-test (*β* = 0.06, *SE* = 0.11, *z* = 0.51, *p* = 0.61).

Across the three groups, participants' performance increased significantly after training (Mpre = 0.70, SDpre = 0.14, Mpost = 0.76, SDpost = 0.14, *β* = 0.37, *SE* = 0.13, *z* = 2.90, *p* < .01). There was no significant difference in the improvement for trained and untrained items (word-type by test-session interaction: *β* = 0.08, *SE* = 0.16, *z* = 0.49 *p* = .63). The interactions between the variability contrasts and test-session were not significant (LV versus HV: *β* = -0.20, *SE* = 0.31, *z* = -0.65, *p* = .52; LV versus HVB: *β* = -0.31, *SE* = 0.31, *z* = -0.99, *p* = .32), and they were not qualified by any higher level interactions with *item-novelty* (LV versus HV: *β* = 0.01, *SE* = 0.38, *z* = 0.02, *p* = .99; LV versus HVB: *β* = -0.30, *SE* = 0.38, *z* = -0.79, *p* = .44).

3.3.1.3 Pinyin accuracy

The predicted variable was whether the participants produced the correct string of phonemes (1/0) in each trial (as determined by the rater). The predictors were *test-session* (pre-test, post-test), *variability-condition* (LV versus HV, LV versus HVB) and *item-novelty* (trained, untrained). Mean pinyin accuracy is displayed in *Figure 3.3.1.3*.

At pre-test, there was no significant difference between the LV and the HV group (*β* = -0.05, *SE* = 0.13, *z* = -0.41, *p* = .68) nor between the LV and the HVB group (*β* = -0.08, *SE* = 0.13, *z* = -0.60, *p* = .55), suggesting that the groups started at a similar level. However, participants did better on untrained words than trained words at pre-test (*β* = 0.25, *SE* = 0.09, *z* = 2.82, *p* < .01), suggesting potential accidental differences in these items. Participants showed no improvement after training (Mpre = 0.54, SDpre = 0.13, Mpost = 0.55, SDpost = 0.13, *β* = 0.07, *SE* = 0.09, *z* = 0.81, *p* = .42). In addition, there was no evidence of different improvements for different variability conditions (*test-session* by LV versus HV: *β* = -0.02, *SE* = 0.22, *z* = -0.09, *p* = .93; *test-session* by LV versus HVB: *β* = -0.27, *SE* = 0.22, *z* = -1.24 *p* = .22) or any interaction between *variability condition*, *test-session* and *item-novelty* (LV versus HV: *β* = 0.07, *SE* = 0.31, *z* = 0.23, *p* = .82; LV versus HVB: *β* = -0.41, *SE* = 0.31, *z* = -1.33 *p* = .18).

3.3.2 Three Interval Oddity Task

The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were *test-session* (pre-test, post-test), *variability-condition* (LV versus HV, LV versus HVB), *trial-type* (neutral versus easy, neutral versus hard) and *item-novelty* (trained item, untrained item). The mean accuracy is displayed in *Figure 3.3.2*.

At pre-test, there was no significant difference between the LV and the HV group (*β* = -0.002, *SE* = 0.14, *z* = -0.01, *p* = .99) nor between the LV and the HVB group (*β* = 0.12, *SE* = 0.14, *z* = 0.86, *p* = .39), suggesting the groups started at a similar level. However, performance with the items classified as "untrained" was significantly greater at pre-test (*β* = -0.31, *SE* = 0.06, *z* = -4.95, *p* < 0.01), suggesting accidental differences between items. As expected, at pre-test participants performed significantly better on "*easy*" trials (where the target speaker had a different gender) than "*neutral*" trials (where all three speakers had the same gender), *β* = 0.40, *SE* = 0.08, *z* = 5.09, *p* < 0.01; and "*neutral*" trials were marginally easier than "*hard*" trials (where one of the foil speakers had the odd gender out), *β* = -0.14, *SE* = 0.08, *z* = -1.81, *p* = 0.07.

Overall, participants' performance increased significantly after training (Mpre = 0.59, SDpre = 0.21, Mpost = 0.66, SDpost = 0.19, *β* = 0.31, *SE* = 0.05, *z* = 6.54, *p* < .001). Critically, there was no reliable interaction between *test-session* and *item-novelty* (*β* = 0.14, *SE* = 0.09, *z* =

1.49, p = .14), suggesting no evidence that training had a greater effect for trained words than for novel words. There was also no interaction with *test-session* for either the contrast between the LV versus the HV conditions (β = -0.01, SE = 0.12, z = -0.12, p = .90) or the contrast between the LV versus the HVB conditions (β = 0.01, SE = 0.12, z = 0.11, p = .91) and no higher-level interactions. This suggests that the extent to which participants improved on this task between pre and post-test did not differ across *variability-conditions* or *item-novelty*.

Although not part of our key predictions, we also looked to see if there was evidence that participants improved more with the easier or harder trials. In fact, the interaction between *test-session* and the contrast between "*easy*" and "*neutral*" was significant (β = -0.27, SE = 0.11, z = -2.39, p = .02) while the contrast between "*neutral*" and "*hard*" was not (β = 0.12, SE = 0.11, z = 1.06, p = .29). This was due to the fact that there was improvement for "*neutral*" (Mpre = 0.57, SDpre = 0.14, Mpost = 0.65, SDpost = 0.15) and "hard" trials (Mpre = 0.54, SDpre = 0.16, Mpost = 0.65, SDpost = 0.15) but not for "easy" trials (Mpre = 0.66, SDpre = 0.16, Mpost = 0.68, SDpost = 0.15).

3.3.3 Summary of Tests administered Pre-and Post-Training

The analysis of Word Repetition and Three Interval Oddity data showed that participants' performance increased significantly after training (except for Pinyin accuracy in Word Repetition) for both tasks. For the Pinyin accuracy measure in Word Repetition, and for the Three Interval Oddity task, there was a main effect of *item-novelty* at pre-test, suggesting that items designated to be "untrained" were accidentally easier than those designated as "trained", but no interaction with *test-session* suggesting that training did not differentially affect improvement with trained and untrained items. However, the critical finding was that there was no interaction between *test-session* and *variability-condition*, or between *test-session*, *item-novelty* and *variability-condition*, providing no evidence that the variability manipulation affected the extent of improvement in these tests.

3.4 Training Data

Here, a model containing data from all three conditions did not converge; however two separate models, one including the LV and HV conditions, and the other the LV and HVB conditions (with condition as a factor with two levels), did converge. In each case the predicted variable was whether a correct response was given (1/0) on each trial. The predictors were the numeric factor *training-session* (1→6) and the factor *variability-condition* which had two levels (model 1: LV versus HV; model 2, LV versus HVB). The mean accuracy is displayed in *Figure 3.4*.

In both models, there was an effect of *training-session* (model 1: $\beta = 0.49$, $SE = 0.04$, $z = 11.52$, $p < .001$; model 2: $\beta = 0.53$, $SE = 0.04$, $z = 12.17$, $p < .001$): Participants' performance increased significantly with training-sessions. Overall, the LV group performed better than both the HV group ($\beta = -0.79$, $SE = 0.16$, $z = -5.03$, $p < .001$) and the HVB group ($\beta = -0.83$, $SE = 0.32$, $z = -2.61$, $p < .01$). However the LV versus HV contrast was also modulated by an interaction with *test-session* ($\beta = -0.19$, $SE = 0.04$, $z = -4.59$, $p < .001$), as was the LV versus HVB contrast ($\beta = -0.35$, $SE = 0.08$ $z = -4.33$, $p < .001$). From *Figure 3.4* it can be seen that the LV and the HVB group did not differ in the first session (i.e. where they get identical input) but the difference gradually increased over the next sessions. For the LV and the HV group, they differed starting from the first session and this difference continued to increase throughout training.

3.4.1 Summary of training data

The analysis of training data revealed significant improvements for all three groups. The LV group performed better than the other two groups due to repetitive exposure to just one speaker throughout the six sessions. In the first session, the difference between the LV and the HVB groups was not significant. However, the difference between conditions increased over time for both LV-HVB and LV-HV contrasts.

3.5 Tests Administered at Post-Test Only

3.5.1 Picture Identification

The coding and reliability analyses for this data is described in section 3.3.1.1. The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were the factor *voice-novelty* (Trained voice, New voice) and the factor *variability-condition* which had two contrasts (LV versus HV, LV versus HVB). The mean accuracy is displayed in *Figure 3.5.1.1*.

There was a main effect of *voice-novelty* ($\beta = 1.07$, $SE = 0.16$, $z = 6.53$, $p < .001$) reflecting higher performance in trials with trained voices. Participants in the LV group performed better than those in the HV group ($\beta = -0.71$, $SE = 0.32$, $z = -2.23$, $p = .03$) but there was no significant difference between the LV and the HVB group ($\beta = -0.14$, $SE = 0.32$, $z = -0.44$, $p = .66$). There was a significant interaction between voice-novelty and both the LV-HV contrast ($\beta = -1.19$, $SE = 0.35$, $z = -3.43$, $p < .01$) and the LV-HVB contrast ($\beta = -1.11$, $SE =$

0.36, $z = -3.08$, $p < .01$). Breaking this down by condition: for each condition there was significantly better performance with trained than new voices (LV: $\beta = 1.83$, $SE = 0.29$, $z = 6.42$, $p < 0.001$; HV: $\beta = 0.64$, $SE = 0.23$, $z = 2.86$, $p < 0.01$; HVB: $\beta = 0.73$, $SE = 0.26$, $z = 2.82$, $p < 0.01$). Breaking it down by voice-novelty: For new voices, neither of the contrasts between conditions was significant (LV versus HV: $\beta = -0.12$, $SE = 0.26$, $z = -0.45$, $p = 0.65$; LV versus HVB $\beta = 0.41$, $SE = 0.27$, $z = 1.51$, $p = 0.13$). For trained items, there was significantly higher performance in the LV than HV condition, but no difference between the LV and HVB conditions (LV versus HV: $\beta = -1.30$, $SE = 0.44$, $z = -2.97$, $p < 0.01$; LV versus HVB: $\beta = -0.70$, $SE = 0.45$, $z = -1.55$, $p = 0.12$).

3.5.2 Picture Naming

These data used the same two measures as the Word Repetition data (see section 3.4.1), i.e. (i) tone identification accuracy and (ii) pinyin accuracy analysed with two logistic mixed effect models. There was only one predictor, *variability-condition* (LV versus HV, LV versus HVB) for both models. The descriptive statistics are displayed in *Figure 3.5.2*.

For tone accuracy, participants in the LV group performed showed no significant difference compared with the HV group ($\beta = -0.33$ $SE = 0.22$, $z = -1.54$, $p = 0.12$) and the HVB group ($\beta = -0.24$, $SE = 0.22$, $z = -1.13$, $p = .26$). There was also no significant difference between groups in pinyin accuracy (LV versus HV: $\beta = 0.09$, $SE = 0.25$, $z = 0.35$, $p = 0.73$; LV versus HVB: $\beta = -0.04$, $SE = 0.25$, $z = -0.17$, $p = 0.86$).

3.5.3 Summary of Tests Administered at Post-Test Only

In sum, the analysis of the Picture Identification results suggests that on average, participants had higher accuracy on trained voice trials, demonstrating greater ease in identifying the words which had been trained repeatedly with the same speaker. The interaction between *voice-novelty* and *variability-condition* suggests that exclusive training on a single speaker in the LV condition boosted performance specifically for that speaker. Critically, there is no evidence for greater performance with untrained items in either the HV or the HVB condition, in contrast to the hypotheses. For Picture Naming, no significant result was found.

3.6 Analyses with Individual Aptitude

In order to look at the effect of learner aptitude and the interaction between this factor and variability condition, we first calculated the mean accuracy at pre-test on the PCPT for each participant. This was used as a continuous predictor (*aptitude*) and added to each of the models reported above. In addition we added the interaction between this factor and key experimental factors (see *Table 3*). Based on Perrachione et al. (2011) and Sadakata and McQueen (2014), high variability should benefit high aptitude participants only, while low variability would benefit low aptitude participants only. In our design, we used a continuous measure of individual ability rather than a binary division of high and low variability. We therefore predicted a stronger positive correlation between *aptitude* and amount of learning in the high variability condition than in the low variability condition. In the models for the pre- and post-test data (i.e. Three Interval Oddity and Word repetition) this could show up as a three way interaction between *condition, test-session* and *aptitude*. This interaction could possibly be modulated *by item-novelty* (four way interaction), since variability is thought to be key for generalization. In the tests only administered post training, we looked for an interaction between *aptitude* and *condition* (since we have no measure at pre-test, and since there was no novelty manipulation here).

Each model reported in *Table 3* contained all the fixed and random effects included in the original models (although in some cases we had to remove correlations between slopes due to problems with convergence). For each of the new models we first confirmed that adding in the new effects and interactions with the individual measures did not change any of the previously reported patterns of significance for the experimental effects (see script https://osf.io/wdh8a/?view_only=d1557462138447ffbaafaf7a59662df8)[2].

The results are shown in *Table 3*. *Aptitude* can be seen to contribute to the model for training, the Three Interval Oddity task and the pinyin accuracy measure in the Word Repetition and the Picture Identification task; however there was no interaction between *aptitude* and any other factor. Thus there was no evidence that this measure of aptitude correlated with participants ability to benefit from training (no interaction with *test-session*), nor – critically for our hypothesis - did this differ by training condition (no interaction with *condition* or with *test-session* by *condition*).

Although the analyses use a continuous measure of PCPT for the purposes of visualization, *Figure 3.6.1 and Figure 3.6.2* uses the mean accuracy for participants split into aptitude groups using a median split based on their PCPT score. *Figure 3.6.1* demonstrated the

[2] Note that models did not include all of the interactions between aptitude and each of the fixed effects in the original model, due to problems of convergence. Therefore the effects reported in Table 3 are the full set of additional fixed effects included in the new version of the model. For training data, recall that in section 3.4 we could not fit a converging model to the data from all three conditions, and instead presented two models – one for the LV+HV data, one for the LV+HVB data. We therefore attempted to include the effects of *aptitude* in each of these models; however neither model would converge if interactions with training-session were included and so these were removed. In the second model it was also necessary to remove the random slope for training session to achieve convergence.

results for the Three Interval Oddity task and Training task. The post-test data from the Picture Naming and Picture Identification tasks are shown in *Figure 3.6.2.* The production task, Word Repetition is shown in *Figure 3.6.3*. The results of the main effect of aptitude and its interaction with other predictors are summarised in *Table 3*.

In sum, participants with higher aptitude measure were better at the tasks, but there is no evidence either that this affected their improvement due to training, or, critically, their ability to benefit from the different variability exposure sets.

4 Discussion

The current study investigated the effect of different types of phonetic training on English speakers learning of novel Mandarin words and tones. To our knowledge, this is the first study to train naive participants on all four Mandarin tones, using real language stimuli embedded in a word learning task. Learning was examined using a range of perception and production tasks. Following previous literature, we compared three training conditions: low variability (single speaker), high variability (four speakers, presented intermixed) and high variability blocked (four speakers, presented in blocks). We also administered tests designed to tap individual aptitude in the perception of pitch contrasts, adapted from the previous literature. The results indicated that participants' performance increased during training and that training also led to improved performance on pre- to post- tests of discrimination and production, with evidence of generalization to new voices and items. Participants also showed some ability to recall trained words – including their tones – in a naming task administered at post-test. However the only place where we saw any effect of the variability manipulation was in the training task (and with trained items in the picture identification task, which was highly similar to training), where the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

low variability group outperformed both high variability groups. Critically, we found no evidence in any of our tests that high variability input benefitted learning or generalization, nor did we find any evidence of an interaction between individual aptitude and the ability to benefit from high variability training. In the following discussion, we first consider the findings from each task in turn before turning to a more general discussion of our findings concerning the benefit of high variability input.

4.1 Training and Picture Identification Tasks

The training task employed in this study was a 2AFC task, where participants had to identify the correct meaning of a Mandarin word based on its tone. The results from training indicate that participants performed better in the single speaker LV training than in either the multiple speaker HV or HVB groups. This difference was present from the first session for the LV-HV contrast, and from the second session for the LV-HVB contrast (i.e. the first session where the two conditions differ). Greater difficulty with multiple speaker input is line with the findings of Perachione et al. (2011), although the differences did not emerge so rapidly in that study, possibly due to there being fewer trials per session). Intuitively, repeated exposure to the single speaker in the LV condition allows for greater adaption to speaker specific cues, whereas in the HV condition participants have to adapt to multiple speakers. This is particularly difficult in the unblocked HV condition, where trial-by-trial adaption is needed, which is effortful for participants (Magnuson & Nusbaum, 2007). Importantly, however, for all three groups, their performance gradually increased over each session. In combination with the fact that their performance on the other tasks increased after training, this indicates that the training task and materials were effective.

Critically, the *Picture Identification* test– a version of the training task without feedback which was administered post training – replicated this LV benefit for trained items, but demonstrated it did *not* extend to new *untrained speakers.* In fact, performance on *untrained speakers* was similar across conditions: participants performed more poorly than with the *trained* speaker*,* but were nonetheless above chance even with the untrained speaker. This indicates across-speaker generalization which did *not* depend on witnessing speaker variability in training, a point to which we return below.

4.2 Three Interval Oddity Task

Our key test of perceptual discrimination was a three interval oddity task, where participants had to indicate the "different" word from a set of three. In each trial, the two foil words were the same word, and differed from the target word only in tone. Improvement in this task was significant but relatively modest (from 59% to 66%, following 8 training sessions), however there are many aspects which make this task more difficult than those used in previous studies. In particular, having each stimulus produced by a different speaker makes noting the similarity across tokens much harder, something we discovered in pilot work, where even before training participants were near ceiling with an equivalent task speakerin which the same speaker produced all three stimuli within a single trial. This is not a feature of any of the tests used in Perrachione et al. (2011) or Sadakata and McQueen (2014). In addition, we tested all four tone contrasts, including those involving Tone 3 (which Perrachione et al., 2011, did not include since it was considered perceptually the most confusable tone).

It is important to note that since all of the speakers in these test items were new, improvement in this test indicates generalization over speakers. Moreover, we did not see differences in the extent of improvement for *trained* versus *untrained* items, indicating that improved tone discrimination is not item specific. Critically, this improvement following training occurred equally across the three variability conditions, indicating that input variability was not necessary for generalization, a point to which we return below.

Another result from this test was that we found evidence that some trial types were harder than others. Specifically, at pre-test, participants showed greatest performance for trials where one of the speakers was male and the other two were female, and the target "odd man" was the male speaker ("easy" trials). In contrast, they showed worst performance if there was one male and two female speakers, but the "odd man" was one of the female speakers ("hard" trials). Middle level performance was shown for trials where all three speakers were female ("neutral" trials). This is presumably due to participants relying on perceptual cues associated with speaker gender to do the task. Interestingly, our analyses showed that performance only increased for the trials where the odd man was not the lone male (the "neutral" and "hard" ones), and not for those where the male was the odd man. Given that participants are not near ceiling at pre-test (67%), it is perhaps surprising that their trained knowledge of the tone contrasts does not boost their performance. One possibility is although they are now better able to use tone cues, they are also *less* likely to use gender based cues, which they may now realize are less reliable, masking improvement based on tone for these particular test items.

4.3 Production Tasks

In this study, we used two production tasks: a word repetition task, administered pre and post training, and a picture naming task administered at post-test only. In the word repetition task, participants repeated a selection of Mandarin words produced by a native speaker, half of which would occur/had occurred in the training set, and half of which were untrained. We saw a significant, though relatively modest improvement in participants' ability to reproduce the tone of the stimuli, such that it could be identified by a native speaker (from pre- to post- test: 70% to 76%). This provides some evidence that purely perceptual training can influence production, in line with the findings of Bradlow and Pisoni (1999) and Zeromskaite (2014). Moreover the fact that participants showed a small but nevertheless significant increase in their ability to accurately repeat the segmental information (63% to 64% of words produced with correct segments) suggests that even though our training specifically targeted tone discrimination (which was all that was necessary to succeed in the training task) there was some more incidental learning of other aspects of the stimuli. As in the three interval oddity task, we again saw equivalent improvement for both trained and untrained items, and there was no difference in the extent of improvement in the different types of conditions, indicating that transfer did not rely on speaker-variability in the input.

Finally, in our picture naming vocabulary test participants were required to produce the trained words in response to pictures, without prompts. Participants showed some ability to recall both the segmental phonology and the tones, although unsurprisingly, accuracy here was considerably less than in the word repetition test for both (tone accuracy: 47% pinyin accuracy 50%). Again we saw no differences between variability conditions, which is surprising given the substantial literature on vocabulary learning showing that there is a benefit of training with multiple speakers which can be tapped by naming tasks. We return to this point in the following section.

4.4 The Role of High Variability Materials in Training and Generalization

In the current study, across all of our different tests, we did not find either an overall benefit of exposure to high variability training materials, or any interaction between such a benefit and individual aptitude. We consider first the lack of *overall* variability benefit. This finding is in line with the lack of a main effect in the previous tone-training studies, yet it is at odds with some other phonetic training studies (Logan et al. 1991, 1993; Clopper & Pisoni, 2004; Sadakata & McQueen 2013). This suggests the possibility that this overall variability benefit is restricted to segmental rather than tonal phonetic learning, at least for speakers of a non-tonal L1. It is harder to reconcile the lack of benefit for vocabulary learning in the picture naming task, given the findings of Barcroft, Sommers and colleagues (Barcroft & Sommers, 2005, 2014; Sommers & Barcroft, 2007, 2011), particularly for our measure of segmental learning which is quite similar to that used in previous experiments, although the nature of our focused phonetic training is a possible explanation. However, it is also important to acknowledge the limitations of a null result: we have no evidence of an effect, but we also don't have evidence that there is *no* effect (see Dienes, 2008, for discussion of this distinction), and type 2 error is of course a possibility. On the other hand, at least for the phonetic training literature, while there is a longstanding *assumption* that speaker variability is important for generalization, as discussed in the introduction, the original test of this by Logan, Lively and Pisoni was extremely low powered considering they only tested three participants for the learning effect of generalisation. In addition, there are only a handful of published studies which have revisited this result (e.g. Lively et al., 1993; Lee, Perrachione, Dees & Wong, 2007; Gao, Low, Jin & Sweller, 2013). The current results suggest that there is need for further research to establish the extent to which the variability effect is replicable, and the extent to which it applies across different types of linguistic domains.

Turning to the lack of interaction with individual differences, the key question is why our result is different from that of Perrachione et al. (2011) and Sadakata and McQueen (2014). There are a variety of differences across the studies which could underpin the difference. Recall that although we set out to use similar methods to the previous studies, we were unable to use the data from our version of the Sadakata and McQueen test, due to too few participants meeting their inclusion criteria. The test which we did use is similar to that used by Perrachione and Wong, however our task is harder since it uses all six Mandarin vowels (whereas the original study used five, without /*u*/) and all of the Mandarin tones (where they used three, without Tone 3). This change means that that we cannot easily contrast the range of participant scores in the two studies and it may be that the spread of ability of our participant is different from theirs. We also note that our statistical analyses are different from both of the previous studies in that they took their continuous aptitude measures and turned these into binary factors using a “cut off”, where as our statistical approach allows us to use them as continuous variables. However this should in principle make our approach more powerful than in previous studies. Moreover, we included a variety of both perception and production tasks. Thus, even if individual aptitude affects only specific aspects of learning or are only discernible in certain types of tests, we would have expected it to emerge in at least one of our tasks. Again, we have to acknowledge the possibility of type 2 error in our study, particularly since we know that interactions require greater samples than main effects to achieve the same power. On the other hand, type 1 error in the original studies is of course always possible.

4.5 Future Directions

As discussed above, it is difficult to draw strong conclusions from the null effects in the current work. An important limitation here is that – given the differences in materials and tasks compared with previous work - it is not clear what the size of the effects we should have expected. This makes it difficult to conduct a power analysis. It also precludes an informed Bayes factor analysis – which could potentially allow us to differentiate evidence for the null from evidence that is ambiguous (Dienes, 2008) – since this also requires a measure of the predicted effect size for each hypothesis[3]. We therefore suggest that it would be useful to implement a direct, high powered replication of these previous studies. We note that obtaining 90% power would likely require a much larger sample than is standard in these types of studies. Given the time consuming nature of these multiple session training studies, we suggest that moving to online testing may be necessary to make this feasible (see Xie et al. 2018 for an example of an acoustic training study done over the web), or alternately multi-lab collaboration may be necessary.

Although direct replication will play a useful role in establishing these effects, we believe that ultimately it will also be important to develop a more nuanced approach to measuring the factors leading to different levels of aptitude both in tone learning, and in other types of phonetic learning. We note that here in addition to not seeing the predicted interaction with variability, we also didn't see interactions between aptitude and training session in any of our tasks, suggesting that our aptitude measure predicted baseline performance on the task and *not* the ability to improve due to training. In addition, the tasks used to measure "aptitude" are quite similar in nature to the training and test tasks, decreasing their explanatory value. Our ongoing work explores the combined predictive value of a range of measures including measures of attention,

[3] It is possible to inform the H1 using other parts of the same dataset (e.g. see Dienes 2018). However in the current work it was unclear how to do this, particularly for the interactions which are the key hypothesis.

working memory and musical ability. Identifying factors which are predictive of aptitude for tone learning has clear implications for teaching and the personalisation of teaching methods.

5 Conclusion

We trained naive participants on all four Mandarin tones, using real language stimuli embedded in a word learning task. We found improvements in both production and perception of tones which transferred to novel voices and items. We found that learning was greatest for training with a single voice but that training with a single voice versus four voices (whether intermixed or blocked) lead to equal amounts of generalization. Although learner aptitude predicted performance in most tasks, there was no evidence that different levels of aptitude lead to better or worse learning from different types of training input.

References

Aliaga-García, C., & Mora, J. C. (2009). Assessing the effects of phonetic training on L2 sound perception and production. In M. A. Watkins, A. S. Rauber, & B.O. Baptista (Eds.). *Recent Research in Second Language Phonetics/Phonology: Perception and Production* (pp. 2-31). Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Audacity Team. (2015). Audacity (Version 2.1.1). *Computer Program*. Retrieved May, 2015, from http://audacityteam.org/

Alshangiti, W., & Evans, B. G. (2014, May). Investigating the domain-specificity of phonetic training for secondlanguage learning: Comparing the effects of production and perception training on the acquisition of English vowels by Arabic learners of English. In *the Proceedings of the International Seminar for Speech Production, Cologne, Germany*.

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

Bao, Z. (1999). *The Structure of Tone*. New York: Oxford University Press. ISBN 0-19-511880-4.

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

Barcroft, J., & Sommers, M. S. (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/10.1016/j.neuropsychologia.2006.11.015

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

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.14, retrieved 24 July 2015 from http://www.praat.org/

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*, 2074-2085. http://dx.doi.org/10.1121/1.427952

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*, 2299-2310. https://doi.org/10.1121/1.418276

Bygate, M., Swain, M., & Skehan, P. (2013). *Researching pedagogic tasks: Second language learning, teaching, and testing*. London UK: Routledge.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological assessment*, *6*(4), 284. http://dx.doi.org/10.1037/1040-3590.6.4.284

Clopper, C. G., & Pisoni, D. B. (2004). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of phonetics*, *32*(1), 111-140. https://doi.org/10.1016/S0095-4470(03)00009-3

Cohen, A. D. (2014). *Strategies in learning and using a second language*. London UK: Routledge.

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

Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. Psychonomic bulletin & review, 25(1), 207-218. https://doi.org/10.3758/s13423-017-1266-z

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English/I/and/l/accurately. *Language and Speech*, *38*, 25-55. https://doi.org/10.1177/002383099503800102

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*, 649-665. http://dx.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. DOI:10.7717/peerj.3209

Giannakopoulou, A., Uther, M., & Ylinen, S. (2013). Enhanced plasticity in spoken language acquisition for child learners: Evidence from phonetic training studies in child and adult learners of English. *Child Language Teaching and Therapy*, *29*, 201-218. https://doi.org/10.1177/0265659012467473

Iverson, P., Ekanayake, D., Hamann, S., Sennema, A., & Evans, B. G. (2008). Category and perceptual interference in second-language phoneme learning: An examination of English/w/-/v/learning by Sinhala, German, and Dutch speakers. *Journal of Experimental*

Psychology: Human Perception and Performance, *34*, 1305. https://doi.org/10.1037/0096-1523.34.5.1305

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*, 3267-3278. https://doi.org/10.1121/1.2062307

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language*, *59*(4), 434-446. https://doi.org/10.1016/j.jml.2007.11.007

Joanisse, M. F., Manis, F. R., Keating, P., & Seidenberg, M. S. (2000). Language deficits in dyslexic children: Speech perception, phonology, and morphology. *Journal of Experimental Child Psychology*, *77*, 30-60. https://doi.org/10.1006/jecp.1999.2553

Lambacher, S. G., Martens, W. L., Kakehi, K., Marasinghe, C. A., & 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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174. DOI: 10.2307/2529310

Lee, J., Perrachione, T. K., Dees, T. M., & Wong, P. C. (2007, August). Differential effects of stimulus variability and learners' pre-existing pitch perception ability in lexical tone learning by native English speakers. Paper presented at the *16th International Congress of Phonetic Sciences* (pp. 1589-1592).

Lengeris, A., & Hazan, V. (2010). The effect of native vowel processing ability and frequency discrimination acuity on the phonetic training of English vowels for native speakers of Greek. *The Journal of the Acoustical Society of America*, *128*, 3757-3768. https://doi.org/10.1121/1.3506351

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*, 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*, 874-886. https://doi.org/10.1121/1.1894649

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 391. https://doi.org/10.1037/0096-1523.33.2.391

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language*, *65*(2), 145-160. https://doi.org/10.1016/j.jml.2011.04.004

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological methods*, *1*(1), 30. http://dx.doi.org/10.1037/1082-989X.1.1.30

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. *Speech perception, production and linguistic structure*, 113-134.

Perrachione, T. K., Lee, J., Ha, L. Y., & Wong, P. C. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

design. *The Journal of the Acoustical Society of America*, *130*, 461-472. https://doi.org/10.1371/journal.pone.0089642

Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *Cambridge NA Report NA2009/06, University of Cambridge, Cambridge*, 26-46.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413-425. https://doi.org/10.1016/j.jml.2008.02.002

R Development Core Team (2010). R: A Language and Environment for Statistical Computing, Version R 3.3.2. *Available at www. r-project. org. Accessed September*, 2017.

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*, 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,* 1318. https://doi.org/10.3389/fpsyg.2014.01318

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), 231-249. 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), 417-434. https://doi.org/10.1017/S0142716410000469

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

Wang, Y., Jongman, A., & Sereno, J. A. (2003). Acoustic and perceptual evaluation of Mandarin tone productions before and after perceptual training. *The Journal of the Acoustical Society of America*, *113*, 1033-1043. https://doi.org/10.1121/1.1531176

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*, 3649-3658. https://doi.org/10.1121/1.428217

Wong, P., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*, 565-585. https://doi.org/10.1017/S0142716407070312

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, *143*, 2013-2031. https://doi.org/10.1121/1.5027410

Yip, M. (2002). *Tone. Cambridge textbooks in linguistics*. Cambridge: Cambridge University Press.

Zeromskaite, I. (2014). The potential role of music in second language learning: A review article. *Journal of European Psychology Students*, *5,* 78-88. http://doi.org/10.5334/jeps.ci

Table 1 (on next page)

Age mean, age range, average number of language learned and mean starting age of learning the first L2 for participants in each condition.

Condition	Age Mean	Age Range	Languages Learned	Average Staring Age
Low Variability	26.15	19-53	2.7	13.8
High Variability	25.65	19-47	2.5	12.2
High Variability Blocking	22.05	19-30	2.0	11.8

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

Table 2 (on next page)

Counterbalancing of voices for each task, training condition and version. LV = Low Variability; HV = High Variability; HVB = High Variability Blocking; PCPT = Pitch Contour Perception Test; CSTC = Categorisation of Synthesized Tonal Continua.

Task	Condition	Voice				
		Version 1	*Version 2*	*Version 3*	*Version 4*	*Version 5*
Training	LV	F1	F2	F3	M1	M2
	HV & HVB	F1	F2	F3	M1	M2
		F3	F1	M2	F1	F2
		M1	M1	F1	F2	F3
		M2	M2	F2	F3	M1
Word Repetition	All	F1	F2	F3	M1	M2
Picture Identification						
Trained Items	All	F1	F2	F3	M1	M2
New Items	All	F2	F3	M1	M2	F1
Three Interval Oddity	All	All versions: MN1, FN1, FN2, FN3				
PCPT	All	All versions: MN1, FN1, FN2, FN3				
CSTC	All	All versions: Synthesized voice				

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

Table 3 (on next page)

Statistics obtained when adding in participant aptitude (as measured by performance on the Pitch Contour Perception Test task at pre-test) into the models predicting performance on the test and training tasks.

1

Data Set	Coefficient Name	Statistics
Word Repetition: Tone Accuracy (Pre/Post)	Aptitude	β = 0.28, SE = 0.42, z = 0.68, p = .496
	Aptitude by *Test-Session*	β = -0.56, SE = 0.71, z = -0.79, p = .429
	Aptitude by LV-HV Contrast by *Test-Session*	β = 0.96, SE = 1.77, z = 0.54, p = .587
	Aptitude by LV-HVB Contrast by *Test-Session*	β = 0.11, SE = 1.51, z = 0.07, p = .941
	Aptitude by LV-HV Contrast by *Test-Session* by *Item-Novelty*	β = -0.84, SE = 2.01, z = -0.42, p = .676
	Aptitude by LV-HVB Contrast by *Test-Session* by *Item-Novelty*	β = 0.29, SE = 1.78, z = 0.16, p = .872
Word Repetition: Pinyin Accuracy (Pre/Post)	**Aptitude**	**β = 0.62, SE = 0.27, z = 2.31, p = .021**
	Aptitude by *Test-Session*	β = -0.28, SE = 0.51, z = -0.56, p = .576
	Aptitude by LV-HV Contrast by *Test-Session*	β = -0.07, SE = 1.28, z = -0.05, p = .958
	Aptitude by LV-HVB Contrast by *Test-Session*	β = -0.57, SE = 1.10, z = -0.52, p = .602
	Aptitude by LV-HV Contrast by *Test-Session* by *Item-Novelty*	β = -1.70, SE = 1.74, z = -0.98, p = .328
	Aptitude by LV-HVB Contrast by *Test-Session* by *Item-Novelty*	β = 0.21, SE = 1.55, z = 0.14, p = .892
Three Interval Oddity (Pre/Post)	**Aptitude**	**β = 0.68, SE = 0.31, z = 2.19, p = .029**
	Aptitude by *Test-Session*	β = 0.08, SE = 0.27, z = 0.31, p = .757
	Aptitude by LV-HV Contrast by *Test-Session*	β = 0.51, SE = 0.67, z = 0.77, p = .443
	Aptitude by LV-HVB Contrast by *Test-Session*	β = 0.48, SE = 0.58, z = 0.83, p = .409
	Aptitude by LV-HV Contrast by *Test-Session* by *Item-Novelty*	β = 1.20, SE = 1.28, z = 0.94, p = .345
	Aptitude by LV-HVB Contrast by *Test-Session* by *Item-Novelty*	β = -0.60, SE = 1.14, z = -0.52, p = .602
Training	**Aptitude**	**β = 0.91, SE = 0.31, z = 2.93, p = .003**

(Model including LV and HV conditions and LV only)	Aptitude by LV-HV Contrast	β = -0.43, SE = 0.33, z = -1.31, p = .192
Picture Identification (Post Only)	**Aptitude**	**β = 1.48, SE = 0.75, z = 1.97, p = .049**
	Aptitude by Voice Novelty	β = -0.28, SE = 0.86, z = -0.33, p = .744
	Aptitude by LV-HV Contrast	β = -0.23, SE = 1.85, z = -0.13, p = .899
	Aptitude by LV-HVB Contrast	β = 0.14, SE = 1.63, z = 0.09, p = .931
	Aptitude by LV-HV Contrast by *Voice-Novelty*	β = 3.47, SE = 2.07, z = 1.68, p = .094
	Aptitude by LV-HVB Contrast by *Voice-Novelty*	β = -1.07, SE = 1.82, z = -0.59, p = .558
Picture Naming: Tone Accuracy	Aptitude	β = 0.38, SE = 0.50, z = 0.75, p = .452
	Aptitude by LV-HV Contrast	β = -0.89, SE = 1.25, z = -0.71, p = .478
	Aptitude by LV-HVB Contrast	β = 0.11, SE = 1.09, z = 0.10, p = .921
Picture Naming: Pinyin Accuracy	**Aptitude**	**β = -1.09, SE = 0.56, z = -1.93, p = .053**
	Aptitude by LV-HV Contrast	β = 0.09, SE = 1.41, z = 0.06, p = .950
	Aptitude by LV-HVB Contrast	β = -0.10, SE = 1.23, z = -0.08, p = .939

2

Figure 1

Tasks completed in each of the eight sessions. (PCPT = Pitch Contour Perception Test; CSTC = Categorisation of Synthesized Tonal Continua).

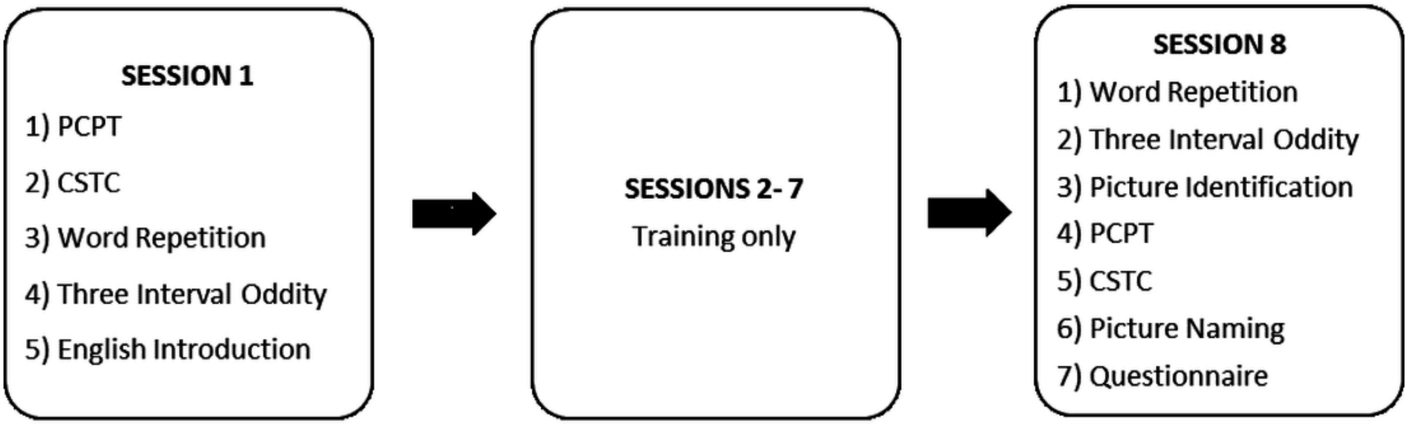


Figure 2

Screen shot from the training task. The stimuli heard is 'dì', tone 4, [earth]. The foil picture on the right is 'dí' tone 2, [siren].

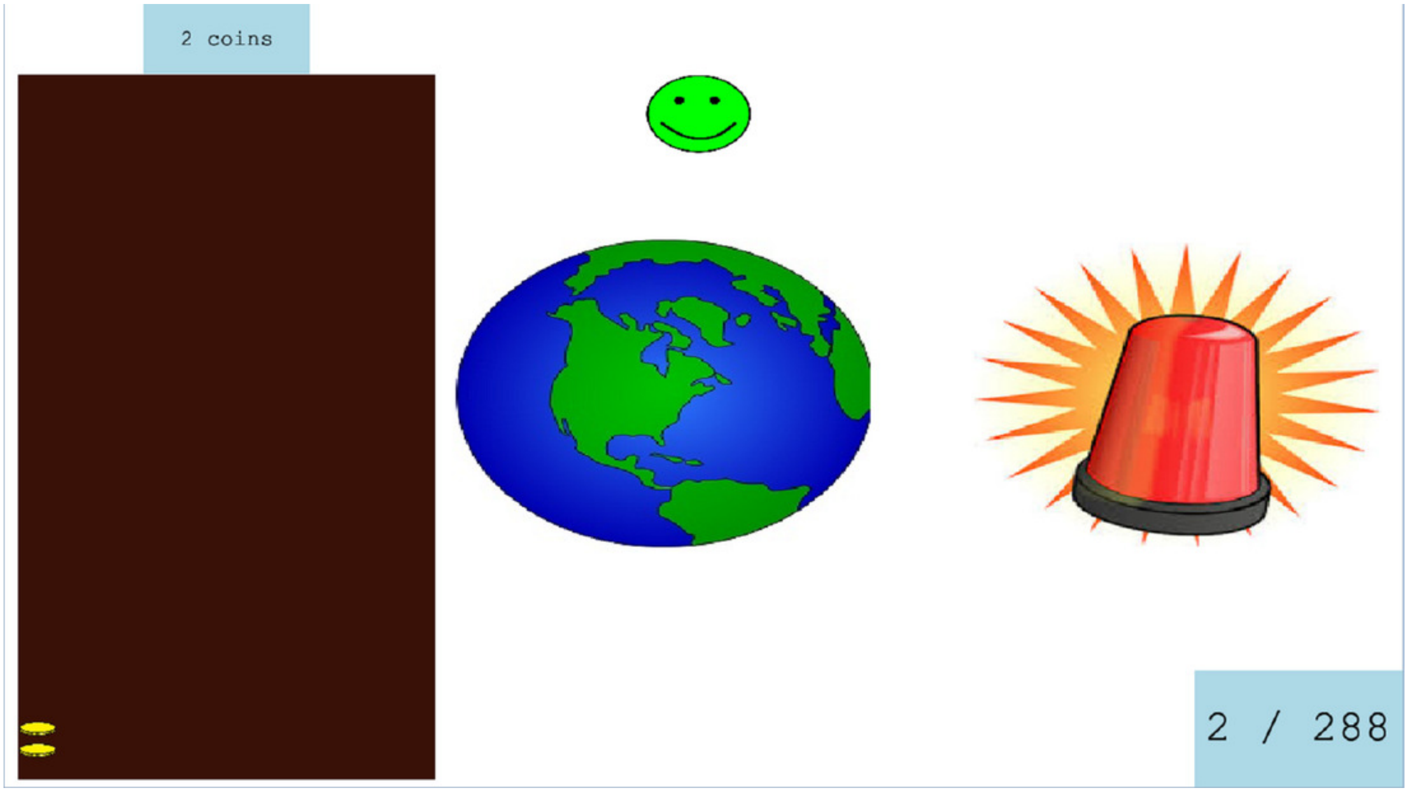


Figure 3

Mean Accuracy from LV (Low Variability), HV (High Variability) & HVB (High Variability Blocking) groups in Pitch Contour Perception Test. Error bars represents the 95% confidence intervals.

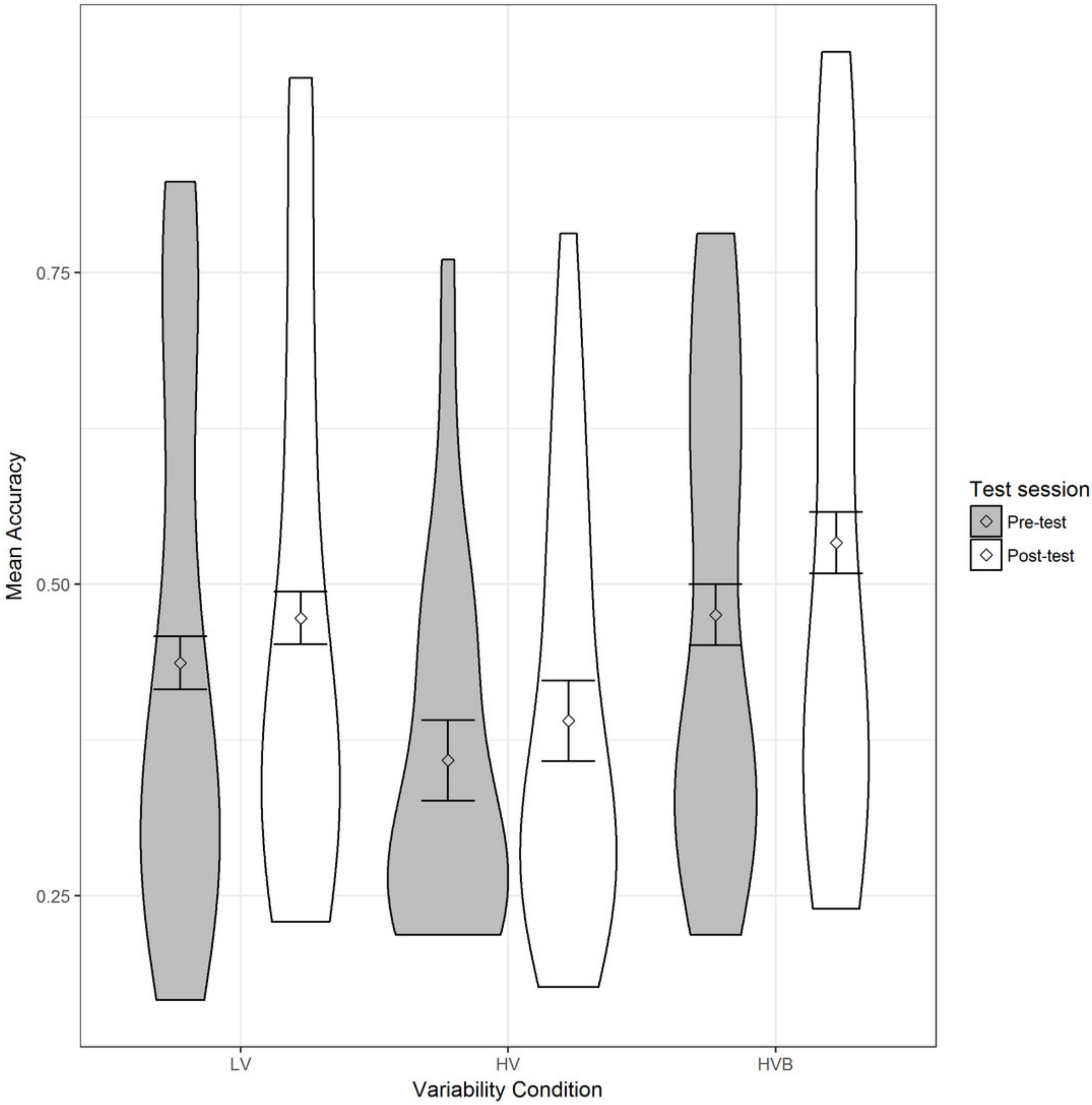


Figure 4

Accuracy of Word Repetition for LV (Low Variability), High Variability (HV) and High Variability Blocking (HVB) training groups in Pre- and Post-tests. Error bars show 95% confidence intervals.

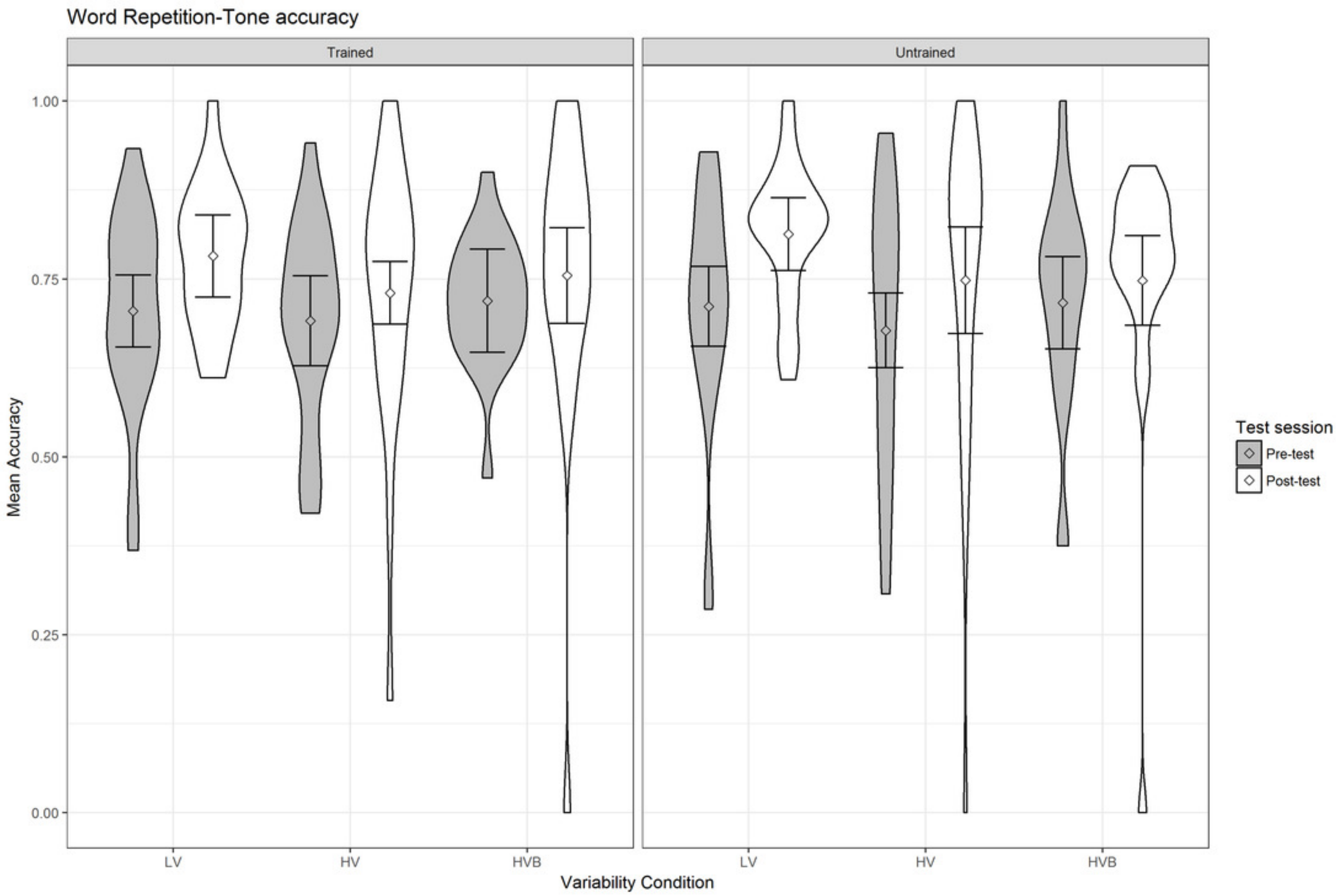


Figure 5

Mean pinyin accuracy of Word Repetition for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups in Pre- and Post-tests. Error bars show 95% confidence intervals.

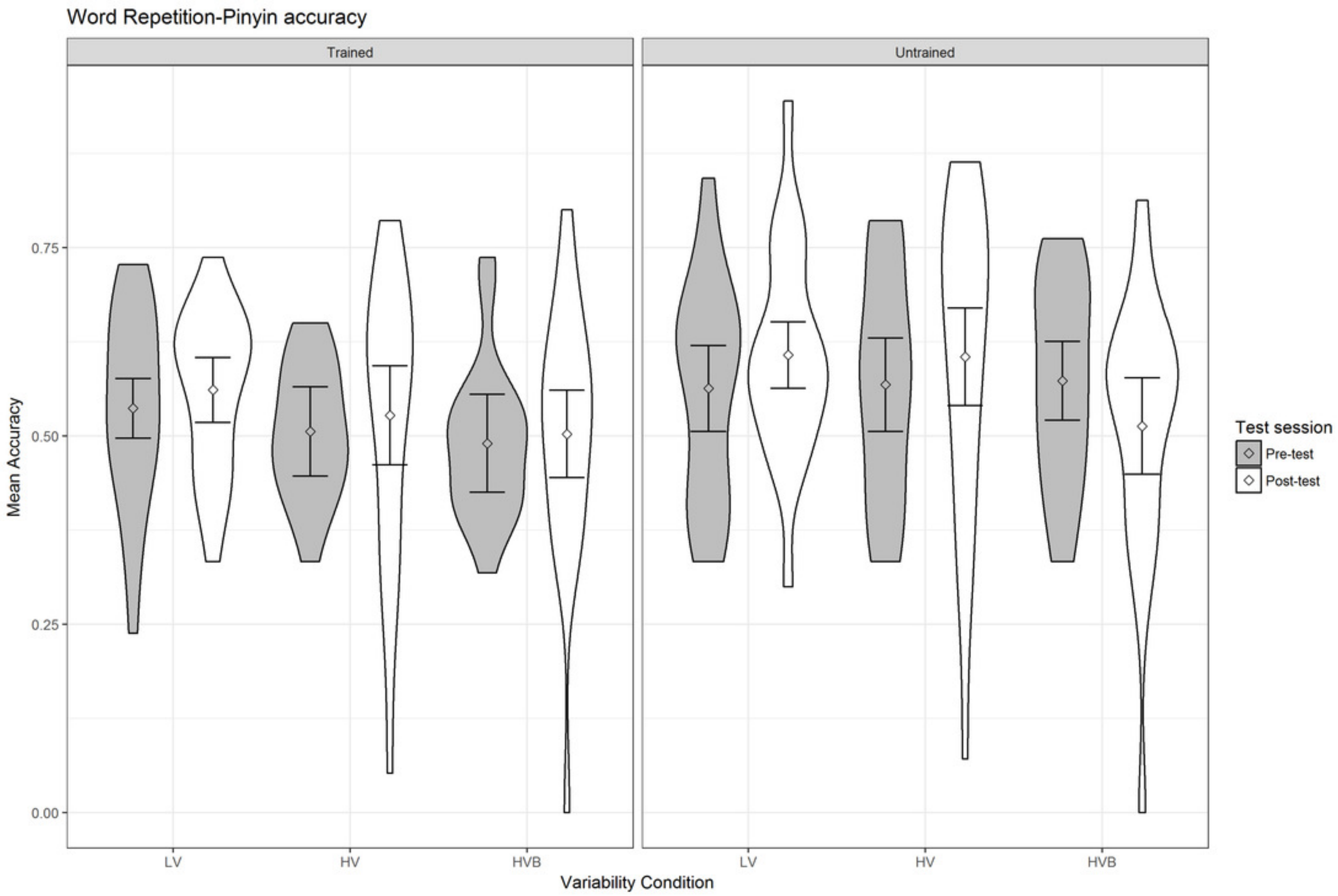


Figure 6

Mean accuracy in Three Interval Oddity task for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups in Pre- and Post-tests. Error bars show 95% confidence intervals.

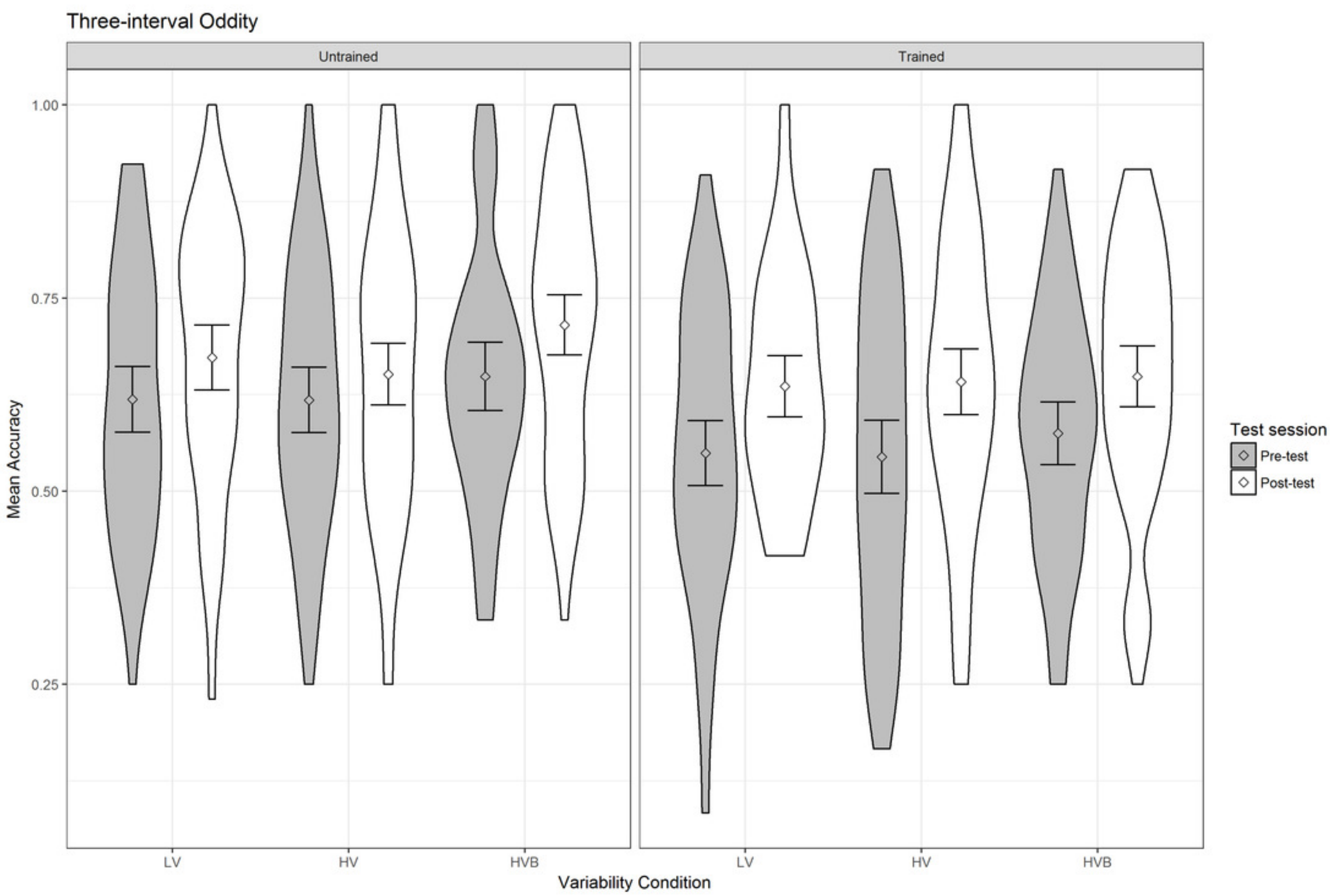


Figure 7

Mean accuracy of Training for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups for each session. Y-axis starting from chance level. Error bars show 95% confidence intervals.

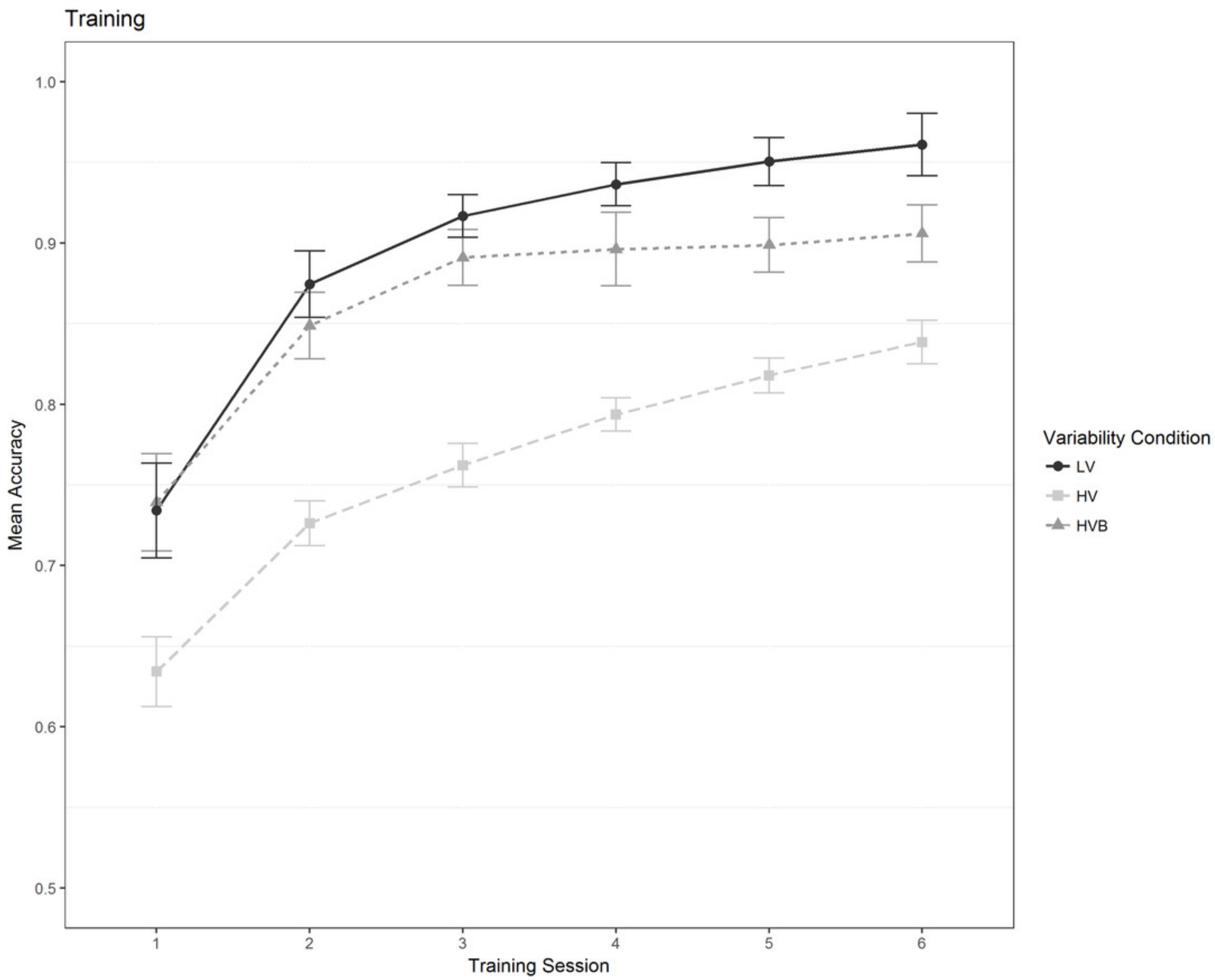
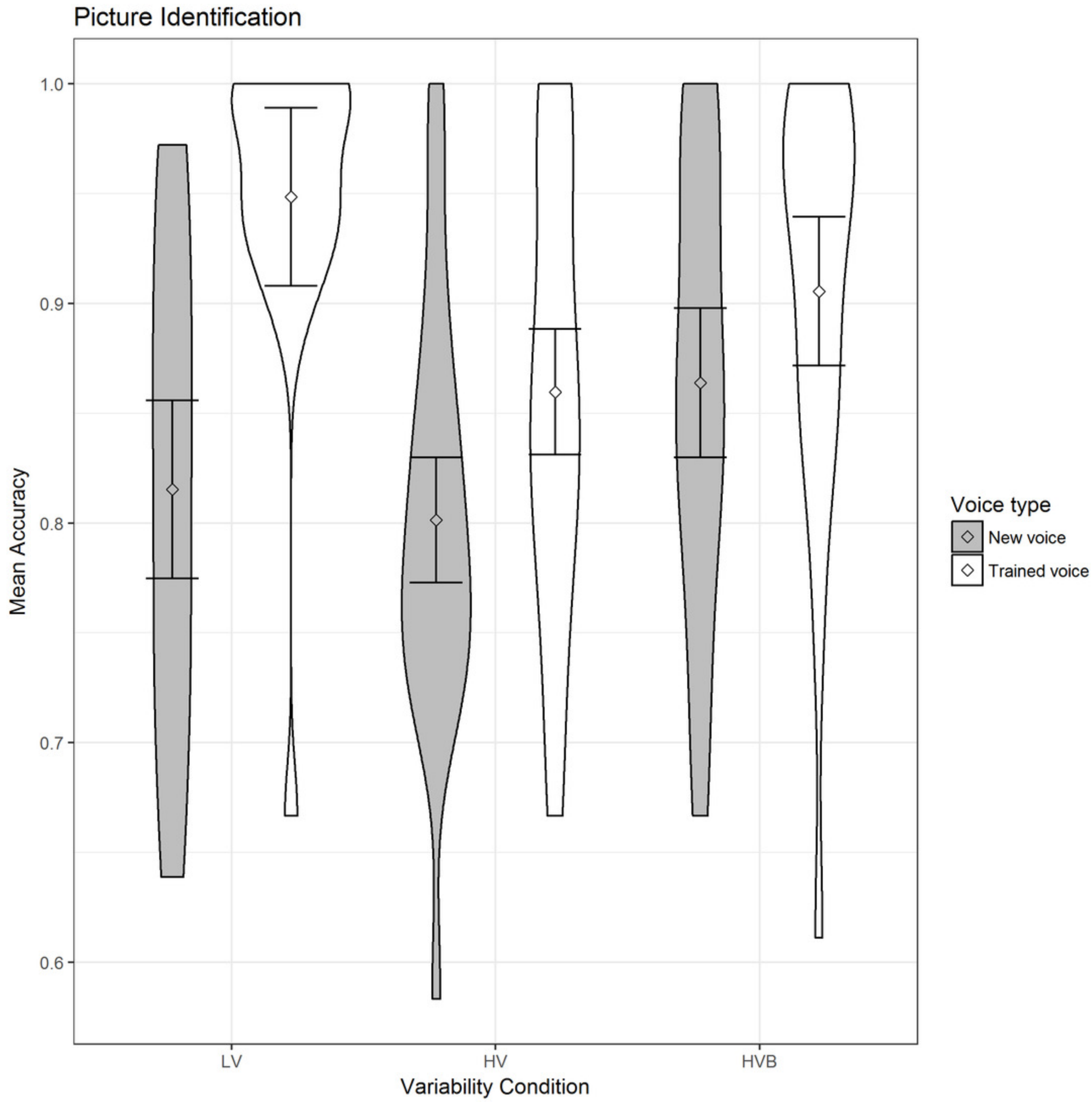


Figure 8

Mean accuracy of Picture Identification for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups for new voices and trained voices. Error bars show 95% confidence intervals.



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27063v1 | CC BY 4.0 Open Access | rec: 29 Jul 2018, publ: 29 Jul 2018

Figure 9

Mean tone accuracy and pinyin accuracy of Picture Naming for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups. Error bars show 95% confidence intervals.

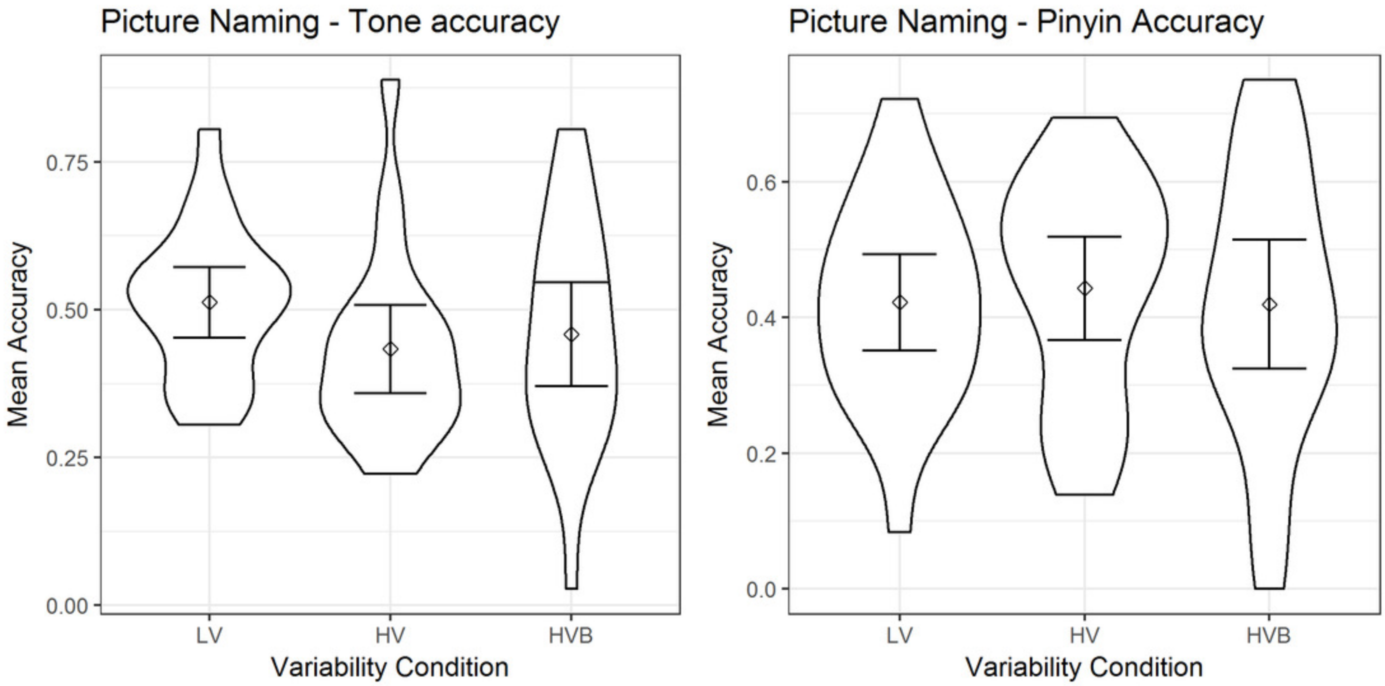


Figure 10

Violin plot for Tone accuracy and Pinyin accuracy of Picture Naming for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups. Error bars show 95% confidence intervals.

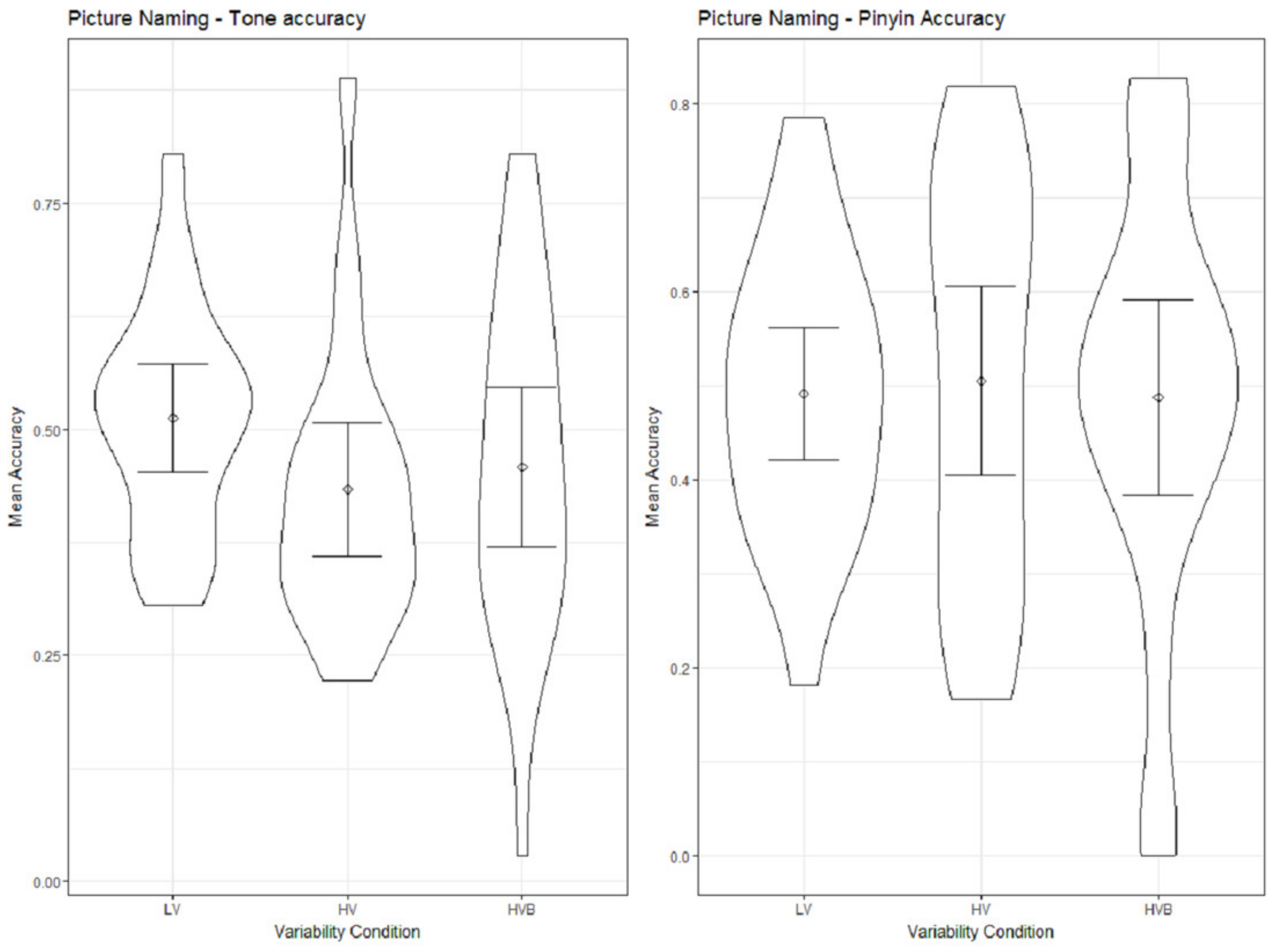


Figure 11

[i]Accuracy in the Three Interval Oddity and Training data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups, split by high versus low aptitude in the PCPT task. Error bars show 95% confidence intervals.[i

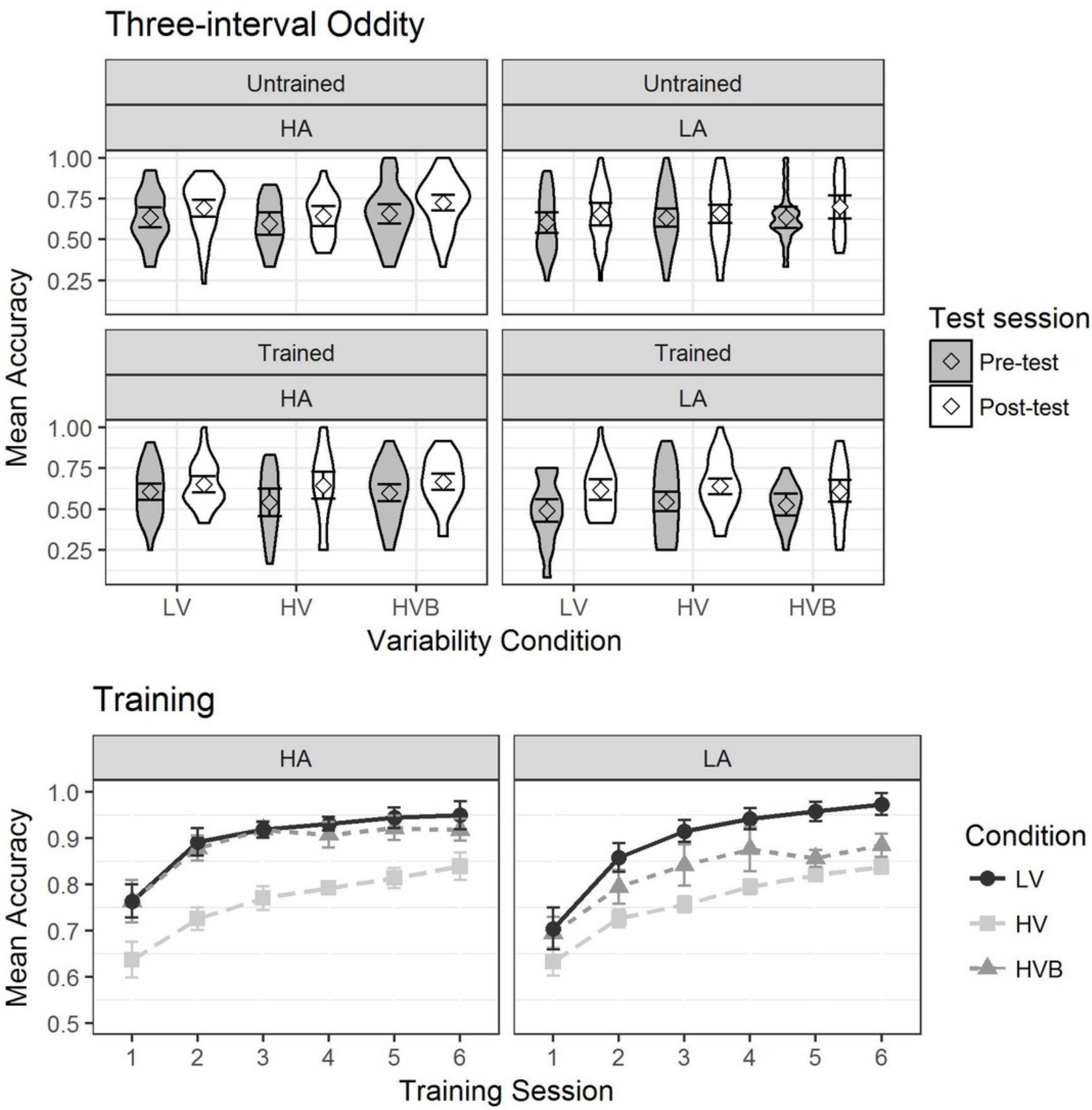


Figure 12

[i]Accuracy in the Picture Naming and Picture Identification data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups, split by high versus low aptitude in the PCPT task. Error bars show 95% confidence inter

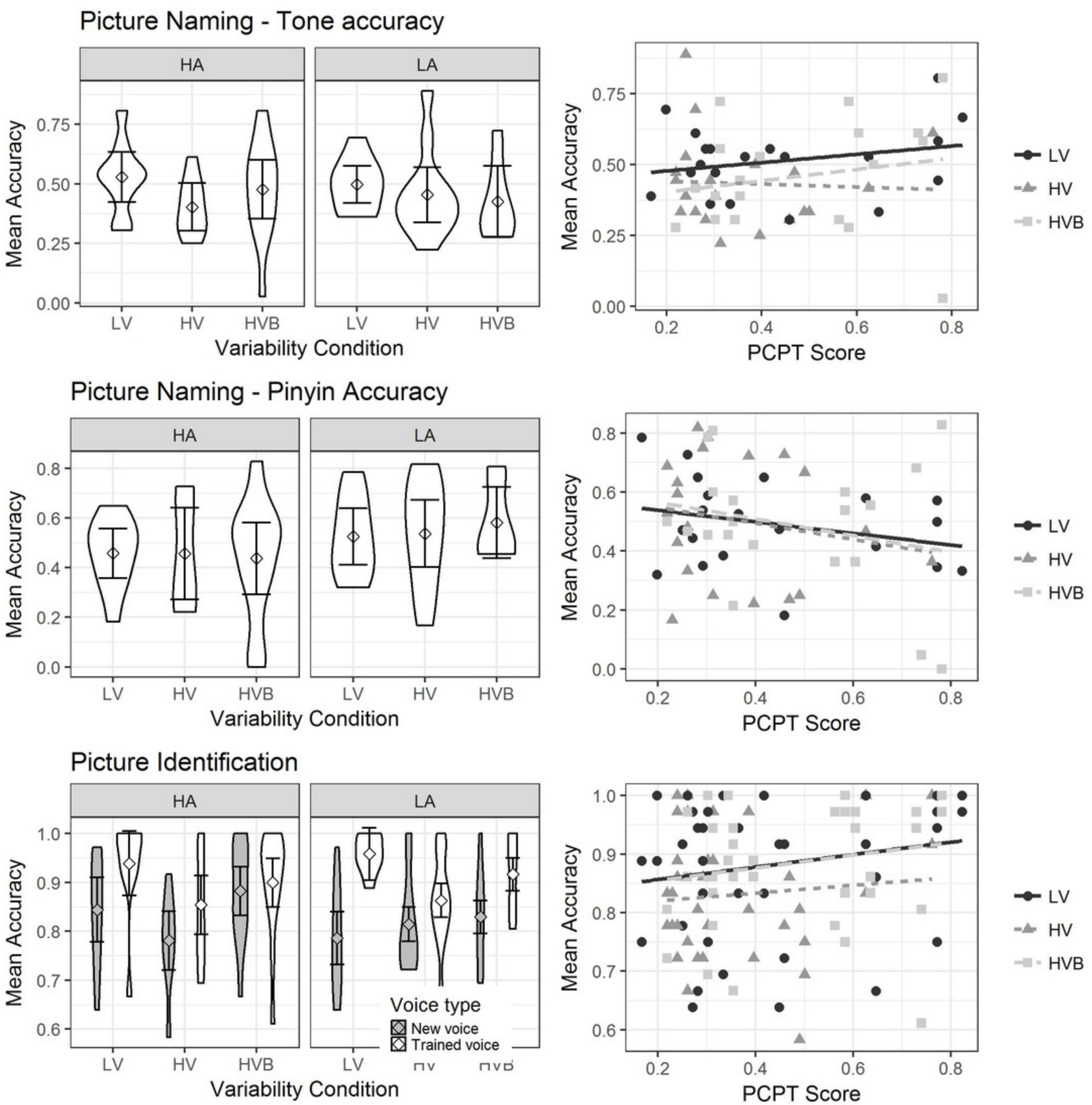


Figure 13

Accuracy in the Word Repetition data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocking) training groups, split by high versus low aptitude in the PCPT task. Error bars show 95% confidence intervals.