

# **The effect of high variability and individual differences on phonetic training of Mandarin tones**

Hanyu Dong

A thesis submitted for the degree of  
Doctor of Philosophy

Division of Psychology and Language Sciences  
University College London

2019

## **Declaration**

I, Hanyu Dong confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

## **Acknowledgements**

First of all, I'd like to give my biggest thank to Elizabeth Wonnacott for being a wonderful supervisor. Your support during my PhD is invaluable – from the detailed guidance on statistical analysis to the insightful advice on all my written pieces and from the swift reply to all my long emails to kind understanding during my difficult times. I've learned so much from you. I'm also very grateful to Bronwen Evans as my second supervisor. Your advice on my research as well as my lifestyle has been a tremendous help.

I'm very thankful for all my fellow language learning lab members: Gwen Brekelmans, Anna Samara, Cat Silvery, Daniela Singh, Maša Vujović, and Elizabeth Wonnacott. I've learned so much from the seminars and workshops provided and organised by these kind lab members. Also, thanks for all the helpful people in Chandler House who has made this building a fantastic place for all people. Special thanks to Merle Mahon, who is very patient and thoughtful to help me with all my personal difficulties and Richard Jardine, who is very helpful regarding all my trivial questions. I'd also like to thank William Webster for providing great support during the build-up of my experiment on Gorilla.sc.

Many thanks to all my fantastic friends: Chen Zhuming, Huang Zhichao, Sunyu Tong, Tao Meiyu, Zhang Shenzhi and Zhu Jingcheng. Thank you very much for all your warm regards and care from different countries. I can't finish the PhD without your help.

Finally and most gratefully, a million thanks to my father, Dong Jun, and my mother Deng Xiaowei. You've been the source of my inner strength and the beacon guiding me through my journey away from home. I can't express enough with my words how grateful and proud I feel to be your son. I also thank all my other family members – thank you for the understanding and support during my PhD!

献给我的姥姥 —— 杨淑芝

献给我的奶奶 —— 廖先桃

## **Abstract**

High variability phonetic training (HVPT) has been found to be more effective than low variability phonetic training (LVPT) in learning various non-native phonetic contrasts. However, little research has considered whether this applies to the learning of tone contrasts. 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). It is also unclear how different types of individual difference measures interact with the learning of tonal language. What work there is, suggests that musical ability is related to discriminating tonal information and in general attention and working memory are linked to language learning. The present study extends these findings by examining the interaction between individual aptitude and input variability and between learning outcomes and individual measures using natural, meaningful L2 input (both previous studies used pseudowords).

In Study 1, forty English speakers took part in an eight-session phonetic training paradigm. They were assigned to high/low variability training groups. High variability used four speakers during the training sessions while low variability used one. All participants learned real Mandarin tones and words. Individual aptitude was measured using an identification and a categorisation task. Learning was measured using a categorical discrimination task, an identification task and two production tasks. 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 of the training material compared with previous research. Although the low variability group exhibited better learning during training than the high variability group, there was no evidence that the different variability training conditions led to different

performances in any of the tests of generalisation. Moreover, although performance on one of the aptitude tasks significantly predicted overall performance in categorical discrimination, identification and training tasks, it did not predict improvement from pre- to post- test. Critically, there was also no interaction between individual aptitude and variability-condition, contradicting with previous findings.

One possibility was that the high variability condition was too difficult as speakers were randomly presented during training, resulting in low trial-by-trial consistency. This greater difficulty might block any advantage of variability for generalisation. In order to examine this, Study 2 recruited additional 20 native English speakers and tested them in a further condition, identical to the previous high variability condition except that each speaker was presented in their own block during the training. Although participants performed better in training compared with the high variability group from study 1, there was again no difference in generalisation compared with the previous conditions, and again no interaction between individual aptitude and variability-condition was found. Bayes Factors were also used to assess the null results. There was evidence for the null for the benefits of high variability for generalisation but only ambiguous evidence regarding whether there was interaction between variability and individual aptitude.

The HPVT used in Study 1 and Study 2 did not replicate the interaction between variability-condition and aptitude found in previous studies. Moreover, although one of the measures of aptitude did correlate with the baseline measures of performance, there was no evidence that it predicted *learning* due to training. Additionally, the two individual aptitude measures used in Study 1 and 2 – taken from Perrachione, et al. (2011) and Sadakata and McQueen (2013) – are not comprehensive. They are natural language-related tasks which directly measure tone perception itself, rather than the underlying cognitive

factors which could underpin this ability. Another interesting question is whether these different cognitive factors might contribute to learners at different stages differently, particularly since language training studies vary as to whether they use current learners of the language or naïve participants, a factor may contribute towards differing findings in the literature.

To explore these issues, Study 3 investigated the relationship between a battery of cognitive individual difference measures and Mandarin tone learning. Sixty native English speakers (forty of whom were currently studying Mandarin at undergraduate level, twenty of whom were naïve learners) took part in a six-session training paradigm. With high-variability training stimuli similar to that used in Study 2 (four speakers blocked), their learning outcomes were assessed by identification, categorical discrimination and production tasks similar to Study 1. Their working memory, attention and musical ability were also measured. Overall, both groups showed improvements during training and in the generalisation tasks. Although Mandarin learner participants performed better than naïve participants overall, the improvements were not generally greater than naïve participants. Each of the individual difference measures was used to predict participant's performance at pre-test and their improvement due to training. Bayes Factors were used as the key method of inference. For Mandarin learner participants, both performances at pre-test and pre- to- post improvement were strongly predicted by attention measures while for naïve speakers, musical ability was the dominant predictor for pre- to- post improvement.

This series of studies demonstrates that Mandarin lexical tones can be trained using natural stimuli embedded in a word learning task and learning generalises to untrained voices and items as well as to production. Although there is no evidence in the current data that the type of training materials affected learning outcomes, tone learning is indeed

affected by individual cognitive factors, such as attention and musical ability, with these playing a different role for learners at different stages.

## Impact statement

Second language learning is increasingly important in the 21<sup>st</sup> century. The process of globalisation has accelerated the cooperation between the Western and Eastern countries. While learning English is compulsory in many Eastern countries such as China, the learning of Mandarin, the official Chinese language, has also received increasing attention. One of the greatest areas of difficulty for European language users learning Mandarin is the learning of the system of lexical tones. The current thesis explores whether native English speakers can learn Mandarin lexical tones using computerized training. Different types of training are compared using high variability (where learners hear multiple speakers exemplifying the tones) and low variability (where learners hear a single multiple speaker exemplifying the tones) training materials. It also investigates how cognitive individual differences including measures of working memory, attention and musical ability contribute to the learning process for learners at different stages (naïve learners with no experience of any tonal language versus individuals who are currently learning the target language at university level). The current research sheds light on how variability and individual differences interact and how learners at different stages benefit differently from individual differences.

The thesis also makes a methodological contribution by introducing the use of Bayes Factors into the field of phonetic training. Bayes Factor analysis is used in Study 2 as a means to evaluate the evidence for the null for non-significant results. In Study 3, it is used as the main method of inference for evaluating the role of the cognitive predictors. Bayes Factor analysis allows us to determine the extent to which results provide evidence for the experiment hypothesis, the null hypothesis or there is only ambiguous evidence. This is in contrast to the traditional frequentist approach using *p*-values, which cannot distinguish evidence for the null from ambiguous evidence.

There are various practical implications from these results. First, they suggest that it is possible to improve perception and identification of Mandarin lexical tones in English speakers using computerised phonetic training. This training can include all four Mandarin tones embedded in real Mandarin stimuli. The results also indicate that learning from a single speaker may be sufficient to learn at least the basic tonal differences in Mandarin. However, we should remain cautious and not overgeneralise this finding as it remains to be seen whether there are other contexts in which hearing multiple speakers is more important (e.g. in connected speech). Cognitive factors are also found to affect the learning of Mandarin tones, and this may differ at different stages of learning. This should be taken into consideration when design training materials and an important line of future research is to examine which type of material may be more efficient for learners with different cognitive profiles.

## Table of contents

Declaration .....	2
Acknowledgements .....	3
-Abstract .....	5
Impact statement .....	9
Table of contents .....	11
List of tables .....	16
List of figures .....	20
<b>1. General introduction .....</b>	<b>26</b>
<b>1.1 Models of L2 speech perception .....</b>	<b>28</b>
<b>1.2 The phonetics of Mandarin .....</b>	<b>30</b>
<b>1.3 Perception of Mandarin Tones by Native English speakers .....</b>	<b>36</b>
<b>1.4 High Variability Phonetic Training .....</b>	<b>39</b>
1.4.1 Phonetic Training of Non-Tonal Contrasts .....	39
1.4.2 Phonetic Training of L2 Lexical Tones .....	42
<b>1.5 Overview of the current thesis .....</b>	<b>48</b>
<b>2. Study 1 .....</b>	<b>52</b>
2.1 Introduction .....	52
2.1.1 The studies by Perrachione et al., 2011 and Sadakata & McQueen, 2014 .....	52
2.1.2 The current study .....	56

2.2	Method .....	59
2.2.1	<i>Participants</i> .....	59
2.2.2	<i>Stimuli</i> .....	60
2.2.3	<i>Procedure</i> .....	63
2.3	Results .....	69
2.3.1	<i>Statistical Approach</i> .....	69
2.3.2	<i>Individual Aptitude Tasks</i> .....	72
2.3.3	<i>Training</i> .....	75
2.3.4	<i>Perceptual tests</i> .....	76
2.3.5	<i>Production tests</i> .....	79
2.3.6	<i>Analyses with Individual Aptitude</i> .....	85
2.4	Discussion .....	88
2.4.1	Pitch Contour Perception Test & Categorisation of Synthesised Tonal Continua 89	
2.4.2	Performance in Training and Picture Identification.....	90
2.4.3	Three Interval Oddity Task.....	91
2.4.4	<i>Word Repetition &amp; Picture Naming</i> .....	93
2.4.5	<i>Evaluation</i> .....	94
<b>3.</b>	<b>Study 2</b> .....	<b>97</b>
<b>3.1</b>	<b><i>Introduction</i></b> .....	<b>97</b>
<b>3.2</b>	<b><i>Method</i></b> .....	<b>98</b>
3.2.1	Participants.....	98

3.2.2	Stimuli & Procedure .....	99
3.3	Results .....	100
3.3.1	<i>Statistical Approach</i> .....	100
3.3.2	<i>Individual Aptitude Tasks</i> .....	101
3.3.3	<i>Training</i> .....	104
3.3.4	<i>Perceptual tests</i> .....	105
3.3.5	<i>Production tests</i> .....	109
3.3.6	<i>Analyses with Individual Aptitude</i> .....	114
3.3.7	<i>Bayes Factor Analyses</i> .....	123
<b>3.4</b>	<b><i>Discussion</i></b> .....	<b>130</b>
3.4.1	Results of perception tasks.....	131
3.4.2	Results of individual aptitude measures .....	134
3.4.3	<i>Limitations and Future directions</i> .....	135
<b>4.</b>	<b>Study 3</b> .....	<b>137</b>
<b>4.1</b>	<b><i>Introduction</i></b> .....	<b>137</b>
4.1.1	Working memory .....	138
4.1.2	<i>Attention</i> .....	148
4.1.3	<i>Musical ability</i> .....	159
<b>4.2</b>	<b><i>The current study</i></b> .....	<b>165</b>
4.2.1	<i>Measures of individual aptitude</i> .....	167
4.2.2	<i>Training and testing paradigm</i> .....	171

4.3	Method .....	174
4.3.1	Participants.....	174
4.3.2	<i>Stimuli</i> .....	175
4.3.3	<i>Procedure</i> .....	176
<b>4.4</b>	<b>Results</b> .....	180
4.4.1	<i>Comparison of the participant groups for each of the individual aptitude tasks</i> 181	
4.4.2	<i>Performances measures: Tests Administered Pre- and Post- Training</i> .....	188
4.4.3	<i>Training Data</i> .....	234
<b>4.5</b>	<b>Discussion</b> .....	235
4.5.1	<i>Baseline differences between Naïve Participants and Mandarin Learning Participants</i> .....	237
4.5.2	<i>Relationship between measures of individual differences and performance on the pre- post tests</i> .....	242
4.5.3	<i>Limitations and future directions</i> .....	260
4.5.4	<i>Conclusion</i> .....	265
4.6	Principal component analysis.....	266
4.6.1	Training.....	272
4.6.2	Pinyin Reading.....	273
4.6.3	Four Interval Oddity task.....	276
4.6.4	Pitch Contour Perception Test .....	277
4.6.5	Discussion.....	278
5.	General discussion .....	289

5.1.1	The role of speaker variability in phonetic training.....	292
5.1.2	Factors affecting individual aptitude for tone learning.....	295
5.1.3	Methodological Contribution.....	297
5.1.4	Future research direction.....	299
6.	References.....	304
	Appendix A.....	333
	Trained stimuli:.....	333
	Untrained stimuli .....	339
	Appendix B.....	345
	Appendix C.....	346

## List of tables

<i>Table 1</i> Mandarin finals (vowels) written in Pinyin with IPA form in brackets. Finals marked in read are used in the current experiments. ....	32
<i>Table 2</i> Mandarin initials (consonants) written in Pinyin with IPA form in brackets. Initials marked in read are used in the current experiments. ....	34
<i>Table 3</i> Mandarin Chinese tones.....	35
<i>Table 4.</i> Age mean, age range, average number of language learned and mean starting age of learning the first L2 for participants in each condition.....	60
<i>Table 5</i> Counterbalancing of voices for each task, training condition and version. LV = Low Variability; HV = High Variability; PCPT = Pitch Contour Perception Test; CSTC = Categorisation of Synthesized Tonal Continua.....	63
<i>Table 6</i> 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.....	87
<i>Table 7</i> Statistics obtained when adding in participant aptitude (as measured by performance on the Categorisation of Synthesized Tonal Continua task at pre-test) into the models predicting performance on the test and training tasks. ....	88
<i>Table 8</i> Mean age, age range, average number of languages learned and mean starting age of learning the first L2 for participants in the high variability blocked condition. ....	99
<i>Table 9</i> Counterbalancing of voices for High variability blocked design and Picture Identification. ....	99
<i>Table 10</i> Statistics analysis with Pitch Contour Perception Test as the individual difference measure. ....	116
<i>Table 11</i> Statistics analysis with Categorisation of Synthesized Tonal Continua as the individual difference measure.....	117

<i>Table 12</i> Bayesian analysis for Picture Identification, Picture Naming, Word Repetition and Three Interval Oddity, with red cells representing evidence for the Null and yellow cells representing ambiguous results.....	126
<i>Table 13</i> Bayesian analysis with PCPT as the ID measure, with red cells representing evidence for the Null and yellow cells representing ambiguous results.....	129
<i>Table 14</i> Bayesian analysis with CSTC as the ID measure, with red cells representing evidence for the Null and yellow cells representing ambiguous results.....	130
<i>Table 15:</i> Age mean, age range, average number of language learned and mean starting age of learning the first L2 for participants in each condition. ....	175
<i>Table 16</i> Standardised scores ( $M = 10$ , $SD = 3$ ) for each individual difference measure at pre-test. Numbers in brackets are standard deviations. ....	184
<i>Table 17</i> Regression results and Bayesian factors for individual difference measures at pre-test. Positive $\beta$ indicates larger effect in the MLP group, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results. ....	187
<i>Table 18</i> Regression and Bayesian analysis for Tone accuracy, tone accuracy, with the effect of ID measure and ID measure x participant-condition, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results. ....	196
<i>Table 19</i> Regression and Bayesian analysis for Tone accuracy, tone accuracy, with the effect of ID measure x test-session and ID measure x test-session x participant-condition, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results. ....	197
<i>Table 20</i> Regression and Bayesian analysis for Pinyin reading, pinyin accuracy, with the effect of ID measure and ID measure x participant-condition.....	203
<i>Table 21</i> Regression and Bayesian analysis for Pinyin reading, pinyin accuracy, with the effect of ID measure x test-session and ID measure x test-session x participant-condition,	

*with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results. ....206*

*Table 22 Regression and Bayesian analysis for Four Interval Oddity task, with the effect of ID measure and ID measure x participant-condition, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results. ....209*

*Table 23 Regression and Bayesian analysis for Four Interval Oddity, with the effect of ID measure x test-session and ID measure x test-session x participant-condition, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results. ....215*

*Table 24 Regression and Bayesian analysis for Pitch Contour Perception Test, with the effect of ID measure and ID measure x participant-condition, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results. ....224*

*Table 25 Regression and Bayesian analysis for Pitch Contour Perception Test, with the effect of ID measure x test-session and ID measure x test-session x participant-condition.....230*

*Table 26 Principle Component Analysis of the individual differences measures of Working Memory, Attention and Musical Ability for the naïve participants group. Loadings larger than 0.40 are in bold. ....270*

*Table 27 Principle Component Analysis of the individual differences measures of Working Memory, Attention and Musical Ability for the Mandarin learner participants group. Loadings larger than 0.40 are in bold. ....272*

*Table 28 Regression analysis for Training task with principle components as predictors. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis). ....273*

*Table 29 Regression analysis of Pinyin Reading, tone accuracy with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis). ....274*

Table 30 Regression analysis of Pinyin Reading, pinyin accuracy with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis). .....275

Table 31 Regression analysis of Four Interval Oddity with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis).. .....276

Table 32 Regression analysis of Pitch Contour Perception Test with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis). .....277

Table 33 Summarised results from previous analyses using individual measures as predictors in section 4.4 and current analyses using principal components as predictors (section 4.6.1-4.6.4) for naïve participants. Where cells are grey represents no predictors were identified. ....279

Table 34 Summarised results from previous analyses using individual measures as predictors in section 4.4 and current analyses using principal components as predictors (section 4.6.1-4.6.4) for Mandarin learner participants. Where cells are grey represents no predictors were identified. ....281

## List of figures

<i>Figure 1</i> Mandarin Chinese Lexical Tones (ChinesePod, 2019).....	36
Figure 2 The pitch continuum from T2 to T3 used in Categorisation of Synthesized Tonal Continua.....	62
Figure 3 Tasks completed in each of the eight sessions (PCPT = Pitch Contour Perception Test; CSTC = Categorisation of Synthesized Tonal Continua).....	64
<i>Figure 4</i> Screen shot from the training task. The stimuli heard is 'dì', tone 4, [earth]. The foil picture on the right is 'dī' tone 2, [siren]. .....	68
<i>Figure 5</i> Mean proportion of correct for the LV (Low Variability) & HV (High Variability) groups in Pitch Contour Perception Test. Error bars represents the 95% confidence intervals. ....	72
<i>Figure 6</i> Slope measure for the LV (Low Variability) & HV (High Variability) groups in Pitch Contour Perception Test. Error bars represents the 95% confidence intervals. ....	74
<i>Figure 7</i> Mean proportion of correct in the Training task for the LV (Low Variability) & HV (High Variability) training groups in each session. Y-axis starts from chance level. Error bars show 95% confidence intervals. ....	75
<i>Figure 8</i> Mean proportion of correct in Three Interval Oddity task for LV (Low Variability) and HV (High Variability) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals.....	77
<i>Figure 9</i> Mean proportion of correct of Picture Identification for LV (Low Variability) and HV (High Variability) training groups for untrained voices and trained voices. Error bars show 95% confidence intervals. ....	79
<i>Figure 10</i> Accuracy of Word Repetition for LV (Low Variability) and High Variability (HV) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals. ....	82

*Figure 11* Mean pinyin accuracy of Word Repetition for LV (Low Variability) and HV (High Variability) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals. ....83

*Figure 12* Tone accuracy and Pinyin accuracy of Picture Naming for LV (Low Variability) and HV (High Variability) training groups. Error bars show 95% confidence intervals. ....84

*Figure 13* Mean proportion of correct for the LV (Low Variability), HV (High Variability) & HVB (High Variability Blocked) groups in Pitch Contour Perception Test. Error bars represents the 95% confidence intervals..... 102

*Figure 14* Slope measure for the LV (Low Variability), HV (High Variability) & HVB (High Variability Blocked) groups in the Categorisation of Synthesized Tonal Continua test. Error bars represents the 95% confidence intervals. .... 104

*Figure 15* Mean proportion of correct in the Training task for the LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups in each session. Y-axis starts from chance level. Error bars show 95% confidence intervals..... 105

*Figure 16* Mean proportion of correct in Three Interval Oddity task for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals. .... 107

*Figure 17* Mean proportion of correct of Picture Identification for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups for untrained voices and trained voices. Error bars show 95% confidence intervals. .... 109

*Figure 18* Accuracy of Word Repetition for LV (Low Variability), High Variability (HV) and High Variability Blocked (HVB) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals..... 112

*Figure 19* Mean pinyin accuracy of Word Repetition for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals..... 113

*Figure 20* Tone accuracy and Pinyin accuracy of Picture Naming for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups. Error bars show 95% confidence intervals. .... 114

*Figure 21* Accuracy in the Three Interval Oddity and Training data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, split by high versus low aptitude in the Pitch Contour Perception Test task. Error bars show 95% confidence interval. .... 120

*Figure 22* Accuracy in the Picture Naming and Picture Identification data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, split by high versus low aptitude in the Pitch Contour Perception Test. Error bars show 95% confidence intervals. .... 121

*Figure 23* Accuracy in the Word Repetition data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, split by high versus low aptitude in the Pitch Contour Perception Test task. Error bars show 95% confidence interval ..... 122

*Figure 24* Word Repetition tone accuracy data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, with x-axis as the aptitude in the Categorisation of Synthesized Tonal Continua, and y-axis as the improvement from pre- to post-training. .... 122

*Figure 25* Tasks completed in each of the six sessions. .... 177

*Figure 26* Screen shot from the training task. The stimuli heard is ‘*chuang*’, tone 4. .... 180

*Figure 27* Mean proportion of correct of Digit Span Forward, Digit Span Backward, Letter Number Sequencing and Arithmetic for naïve participants and Mandarin learners. .... 182

*Figure 28* Results of Elevator counting with Distraction, Elevator Counting with Reversal, Telephone Search, Telephone Search while Counting and Visual Elevator for naïve participants and Mandarin learners. .... 183

*Figure 29* Mean proportion of correct of Beat Perception and Melody Memory for naïve participants and Mandarin learners. .... 184

<i>Figure 30</i> Mean tone accuracy of Pinyin reading for naïve participants and Mandarin learners across pre- and post-test.....	196
<i>Figure 31</i> Scatter plot depicting the relationship between participants performance in the Working memory composite score and their improvements from pre- to post- test in the Pinyin reading tone accuracy task.....	200
<i>Figure 32</i> Scatter plot depicting the relationship between participants performance in the Digit span backwards task and their improvements from pre- to post- test in the Pinyin reading tone accuracy task .....	200
<i>Figure 33</i> Mean pinyin accuracy of Pinyin reading for naïve participants and Mandarin learners across pre- and post-test. ....	201
<i>Figure 34</i> Scatter plot for the Pinyin reading pinyin accuracy with working memory composite score as x-axis and pre-test performance as y-axis. ....	204
<i>Figure 35</i> Scatter plot for the Pinyin reading pinyin accuracy with working memory composite score as x-axis and pre-test performance as y-axis. ....	205
<i>Figure 36</i> Mean proportion of correct of Four Interval Oddity task for naïve and learner groups in pre/post sessions,.....	207
<i>Figure 37</i> Scatter plot for the Four Interval Oddity task with PCPT score as x-axis and pre-test performance as y-axis.....	211
<i>Figure 38</i> Scatter plot for the Four Interval Oddity task with LNS score as x-axis and pre-test performance as y-axis. ....	212
<i>Figure 39</i> Scatter plot for the Four Interval Oddity task with Attention composite score as x-axis and pre-test performance as y-axis. ....	212
<i>Figure 40</i> Scatter plot for the Four Interval Oddity task with ECR score as x-axis and pre-test performance as y-axis. ....	213
<i>Figure 41</i> Scatter plot for the Four Interval Oddity task with VE score as x-axis and pre-test performance as y-axis. ....	213

<i>Figure 42</i> Scatter plot for the Four Interval Oddity task with TSC score as x-axis and pre-test performance as y-axis. ....	214
<i>Figure 43</i> Scatter plot for Four-Interval Oddity measure, with PCPT as x-axis and pre/post-test difference as y-axis.....	218
<i>Figure 44</i> Scatter plot for Four-Interval Oddity measure, with Working memory composite score as x-axis and pre/post-test difference as y-axis. ....	218
<i>Figure 45</i> Scatter plot for Four-Interval Oddity measure, with Digit Span Forward score as x-axis and pre/post-test difference as y-axis. ....	219
<i>Figure 46</i> Scatter plot for Four-Interval Oddity measure, with Attention composite score as x-axis and pre/post-test difference as y-axis. ....	219
<i>Figure 47</i> Scatter plot for Four-Interval Oddity measure, with Elevator Counting with Reversal score as x-axis and pre/post-test difference as y-axis. ....	220
<i>Figure 48</i> Scatter plot for Four-Interval Oddity measure, with Telephone Search score as x-axis and pre/post-test difference as y-axis. ....	220
<i>Figure 49</i> Scatter plot for Four-Interval Oddity measure, with Telephone Search while Counting score as x-axis and pre/post-test difference as y-axis. ....	221
<i>Figure 50</i> Scatter plot for Four-Interval Oddity measure, with Melody Memory score as x-axis and pre/post-test difference as y-axis. ....	221
<i>Figure 51</i> Mean proportion of correct of Pitch Contour Perception Test for naïve participants and Mandarin learners across pre- and post-test.....	222
<i>Figure 52</i> Scatter plot for the Pitch Contour Perception Test with LNS score as x-axis and pre-test performance as y-axis. ....	226
<i>Figure 53</i> Scatter plot for the Pitch Contour Perception Test with Attention composite score as x-axis and pre-test performance as y-axis. ....	226
<i>Figure 54</i> Scatter plot for the Pitch Contour Perception Test with ECR score as x-axis and pre-test performance as y-axis. ....	227

*Figure 55* Scatter plot for the Pitch Contour Perception Test with VE score as x-axis and pre-test performance as y-axis.....227

*Figure 56* Scatter plot for the Pitch Contour Perception Test with TS score as x-axis and pre-test performance as y-axis.....228

*Figure 57* Scatter plot for the Pitch Contour Perception Test with TSC score as x-axis and pre-test performance as y-axis. ....228

*Figure 58* Scatter plot for the Pitch Contour Perception Test with BP score as x-axis and pre-test performance as y-axis.....229

*Figure 59* Scatter plot for Pitch Contour Perception Test, with Attention composite score as x-axis and pre/post-test difference as y-axis.....232

*Figure 60* Scatter plot for Pitch Contour Perception Test, with ECR score as x-axis and pre/post-test difference as y-axis. ....232

*Figure 61* Scatter plot for Pitch Contour Perception Test, with TS score as x-axis and pre/post-test difference as y-axis. ....233

*Figure 62* Scatter plot for Pitch Contour Perception Test, with BP score as x-axis and pre/post-test difference as y-axis. ....233

*Figure 63* Mean proportion of correct of Naïve participants and Mandarin learners through session 1 to 4.....234

## 1. General introduction

With the continuing development of globalisation, learning a second language (L2) has become increasingly common. Statistics showed that during year 2017, there were more than 320000 non-EU students coming to UK for education above undergraduate level and 38% among them were Chinese students (UKCISA, 2019). On the other hand, around 492000 international students went to China for a degree in year 2018 (Wang, 2019). Academic communication between Europe and Asia has reached unprecedented levels and learning a language that is largely different from the first language (L1) has become increasingly important. Corresponding to this, while earlier research in second language (L2) learning has mainly focused on European languages, in recent years there has been an increasing amount of research on second language learning outside of this circle.

One key aspect of learning an L2 is to learn the speech sounds it uses, i.e. to learn a new phonological system. There is extensive research suggesting that humans are born with sensitivity to phonological contrasts outside of the native language. Such sensitivity decreases during the first year of life. After that, infants are tuned to the phonetic contrasts of their native language (Werker & Tees, 1984). This is believed to have evolutionary advantages as it allows infants to concentrate on the input which is relevant (Bornstein, Hahn & Haynes, 2004). Due to this early development, learning the sounds of an L2 is extremely difficult for adults. This process can be particularly difficult when the L2 relies on the same acoustic dimensions as the L1, but for different purposes (Bygate, Swain, & Skehan, 2013), suggesting that it is challenging to adjust existing acoustic properties in the L1 to learn new L2 categories.

One difference between the phonological systems of most European languages and many Asian languages is that the latter make use of *lexical tone*. Lexical tone is a type of

phonological contrast whereby the pitch contour is used to distinguish lexical information (Yip, 2002). Currently the most widely spoken language which makes use of lexical tone is Mandarin Chinese (Statista, 2019), which 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. In contrast, non-tonal languages such as English use pitch information extensively for intonation (e.g. forming a question or for emphasis), and although pitch may play a role in marking stress at the lexical level (e.g. *'import/im'port*), this is quite different from a lexical tone system. Thus, learning lexical tone is a key part of learning a language such as Mandarin Chinese and it may be particularly difficult for speakers of language such as English which doesn't involve lexical tones.

The current thesis aims to explore the extent to which *phonetic training* can be used to teach English speakers Mandarin lexical tones. It asks what factors contribute to the effectiveness of training. In particular, it examines how the learning process is affected by two types of factors: the *variability* in training materials and *individual differences* between learners. This general introduction firstly provides background by introducing about common models of L2 speech perception in Section 1.1. Then, Section 1.2 provides a brief introduction to the phonetics of Mandarin, focusing on aspects relevant to the stimuli used in the studies reported the current thesis. Next, Section 1.3 considers how Mandarin lexical tone might cause difficulty for English learners of Mandarin with respect to evidence for the models described in Section 1.1. Section 1.4 focuses on the high variability phonetic training (HVPT) paradigm, first describing how it has been used in training non-tonal phonetic contrasts and the key

findings, followed by how this paradigm has been applied to tone training. Finally, Section 1.5 gives an overview of the structure of the rest of the thesis.

### ***1.1 Models of L2 speech perception***

One of the most important models of L2 speech is the Speech Learning Model (SLM, Flege, 1995). The core of this model suggests that although adults are tuned to their L1 phonetic contrasts, the learning mechanism for the L1 is still intact and is activated when learning an L2. Language-specific aspects of speech sounds are stored in long-term memory representations as *phonetic categories* for both the L1 and L2. Learning an L2 will also create new phonetic categories that don't exist in the L1. Importantly, the degree of success with L2 speech learning is largely determined by the similarity between L1 and L2 sounds. If a new L2 sound is similar to an existing L1 sound, individuals may judge tokens from the L2 category to be from the L1 category, resulting in increased difficulty to perceive and learn the new category. This is named as *phonetic category assimilation*. On the other hand, if an L2 sound differs greatly from any existing L1 sound, individuals will be able to easily identify it as novel, and repeated exposure to L2 sounds will form a new phonetic category, which may lead to *phonetic category dissimilation*. It should be noted that the category dissimilation process can make individuals differentiate the L1 and L2 categories more, i.e. the difference between the L1 and L2 categories from a bilingual speaker will be bigger than it between monolingual speakers of L1 and L2, which has been found with the production studies (Flege, 1987). The SLM highlights the dynamic nature of L2 learning such that with repeated exposure, adult learners can still form new categories parallel to their original L1, while the categories between L1 and L2 may also overlap.

An alternative model of L2 speech learning is the Perceptual Assimilation Model (PAM, Best, 1994; 1995). While SLM focuses on both perception and production, PAM mainly focuses only on the perception process. Another difference is that SLM looks at individual phonetic categories while PAM focuses on pair-wise phonological contrasts. Thus, it provides a much more detailed description of the assimilation mechanism. If an L2 sound is too distant from any L1 category, then it is *non-assimilable* as it cannot be mapped onto any existing category. *Two-category assimilation* results in the most accurate discrimination, where individuals map the L2 contrast onto two different L1 categories. *Single category assimilation* leads to less accurate results since the two members of the L2 contrast are assimilated to the same L1 category and considered as equally good (or poor) members of that category. However there may also be *category goodness assimilation* when the L2 contrast is assimilated into the same L1 category but with one member considered a better member of that category than the other, leading to better discriminability than the *single category assimilation*. It can be seen that similar to SLM, PAM also emphasises that similarity between L2 and L1 may be crucial to L2 learning. Critically, it suggests that the assimilation process is not an “all or nothing” process and the assimilation can be formed in multiple ways as related to the L1 categories.

These two models have received empirical support from many studies. Iverson and Evans (2007) studied how different L1 speakers map English vowels to their L1 vowel system. They recruited German and Norwegian speakers, who have learned a complex vowel system, as well as Spanish and French speakers who have only learned a relatively small vowel system. The results demonstrated that German and Norwegian speakers were more accurate at recognising those English vowels which they were able to map into different L1 vowel categories, while Spanish and French speakers tend to map and assimilate multiple English vowels into the same L1 category. One interpretation is that German and Norwegian speakers show

two category assimilation as suggested by PAM, or dissimilation as suggested in SLM, while Spanish and French show formed single category assimilation due to lack of enough phonetic categories. However, it should also be noted that all the participants did learn new aspects of the English vowel system rather than simply assimilating vowels into existing first-language categories, despite their different learning rates, suggesting that they may have moved beyond the assimilation/dissimilation process. Similar results were reported in a training study in which participants learned the English /w – v/ contrast (Iverson, Ekanayake, Hamann, Sennema, & Evans, 2008). In this study, German and Sinhala speakers, whose L1 contains a single phoneme similar to both English /w/ and /v/, showed slower learning which they suggest is due to single category assimilation as they assimilated both /w/ and /v/ into the same phoneme in their L1. In contrast, Dutch speakers who have two different phonemes in their L1 showed better performance suggesting two category assimilation. It can be seen that previous studies suggest that learning phonetic contrasts can be explained by SLM and PAM. Section 1.3 will look at how these models can be applied to tone learning more specifically. First, however, I discuss the phonetics of Mandarin, a tonal language which is the focus of the current thesis.

## ***1.2 The phonetics of Mandarin***

This section will give a brief overview regarding the phonology of Mandarin. This is important background for the current thesis since all stimuli used in the current training tasks are genuine Mandarin words produced by native Mandarin speakers.

In total, there are more than 80 languages (dialects) in China, more than 60 of which are major languages spoken by more than 1 million speakers (Chinese Academy of Social Sciences, 2012). As a group of languages, Chinese has a relatively complicated background due to the diversity of ethnic groups in China. There are 56 ethnical groups each having a

unique cultural background and a language variations. Among these, Han is the majority ethnical group which contributes to 92% of the Chinese population (National Bureau of Statistics of China, 2010). The language spoken by Han is called HanYu (i.e. Han-Language). Han-Language is divided into seven major dialect groups, Mandarin (Northern Chinese), Wu, Xiang, Gan, Kejia (Hakka), Yue (Cantonese) and Min. Among these, Mandarin speakers comprise around 68% (Chinese Academy of Social Sciences, 2012). In this thesis, “Mandarin” refers to the Modern Standard Mandarin Chinese (MSMC), which is the standard variety of Northern Chinese that is the official language of China. It is mostly widely used in the Mainland China area.

The phonology of Mandarin Chinese can be described in terms of *initials*, *finals*, and *tones*. Most Mandarin syllables consist of one initial (a consonant) followed by one final (vowels or combination of vowels and nasals) with one tone (except for a small proportion of words which only used one final and one tone). Mandarin tones are associated with the entire syllable, rather than a single segment, at the supra-segmental level.

Finals and initials are shown in Tables 1 and Table 2. These tables present the symbols using both IPA and *Pinyin*. Note that standard written Chinese uses a logo-syllabic system where one character generally represents a syllable and each character may be a word or part of a word. *Pinyin* is a parallel, alternative Romanization system which can be used to represent Chinese using a combination of Chinese tone didactics and the Latin alphabet. Pinyin was developed by linguists in the 1950s (Duanmu, 2007) but is now used within the Chinese education system and in teaching Mandarin as an L2. It is designed so that letters correspond roughly to phonemes. The current thesis only involves monosyllabic Mandarin words and - after this chapter - uses Pinyin (rather than IPA) to represent Mandarin words when describing the stimuli. In addition, in Study 3 Pinyin as representations of words are used in the phonetic

training paradigm (see Section 4.3.3.5 and Appendix B). Starting with Table 1, it can be seen that there are 35 finals. These include simple finals comprising a monophthong vowel and more complex finals involving compound vowels or compound vowels followed by either front or back nasals (with difference that back nasals are pronounced with the tongue positioned closer to the palate allowing more air through the nose).

Table 1 Mandarin finals (vowels) written in Pinyin with IPA form in brackets. Finals marked in red are used in the current experiments.

	Main vowels	Compound vowels	Front nasal finals	Back nasal finals
Monophthongs	a[a]		an[an]	ang[aŋ]
	o[ɔ]		en[ən]	eng[əŋ]
	e[ɛ]		in[in]	ing[iŋ]
	i[i]		ün[yn]	ong[uŋ]
	u[u]			
	ü[y]			
Diphthongs		ai[ai]	ian[ien]	iang[jaŋ]
		ao[au]	uan[uən]	uang[uaŋ]
		ei[eɪ]	üan[yeŋ]	ueng[ueŋ] <sup>1</sup>
		ia[ia]	un[uən]	iong[iuŋ]
		ie[ie]		
		ou[ou]		
		ua[ua]		
		üe[ye]		
		uo[uɔ]		
Triphthongs		iao[iau]		
		iu[iou]		
		uai[uai]		
		ui[uei] <sup>2</sup>		

<sup>1</sup> The original spelling was ueng (e.g. 翁/wuēng/, old man) but it was simplified as /wēng/ in written Pinyin. There is no syllable with the initial /w/ and final /eng/ so there is no conflict. In addition, /weng/ is the only syllable involving the back nasal final /ueng/ in Mandarin.

<sup>2</sup> The original spelling was uei (e.g. 会/huei/, meeting) but it was simplified as /hui/ in written Pinyin.

Note that there has been a controversy over the number of main vowels in Mandarin phonology. While the six main vowels shown in the table are used in the education system across the country, linguists argue that this classification is not accurate. For example, Duanmu (2007) argued that there should only be five vowel phonemes in Mandarin: [i], [u], [y], [ə], [a], that the vowel o[ɔ] is just an allophone of the phoneme [ə]. On the other hand, some researchers argue that there are actually more than 6 main vowels, including differences between vowel pronunciations in certain syllables. Cao (2016) suggested that [ê],[ɣ],[-i],[i-] should be added into the system. In Table 1, the finals used in the current study are marked in red. As can be seen, the selection includes a wide range of monophthongs, diphthongs and triphthongs, including both front and back nasals. Among these, vowels involving the novel sound ü[y] (ü[y], ün[yn] & üe[ye]) may be difficult for native English speakers as English does not include these vowels. It is generally agreed that it has 11 non-rhotic distinctive monophthongs /i, ɪ, e, ɛ, æ, ʌ, ɑ, ɔ, o, ʊ, u/. It can be seen that Mandarin contains one vowel missing from the English inventory - the vowel /y/, which is a front rounded vowel. Using the /y/ from German and French, Strange et al., (2007) reported that English speakers found /i-y/ and /u-y/ contrasts particularly hard to discriminate. For the /i-y/ contrast, it is distinguished by lip-rounding, which is a feature doesn't exist in vowel discrimination in English. For the /u-y/ contrast, the two vowels are often assimilated to the same English vowel category /u/ (Levy, 2009). In addition, triphthongs and diphthongs with nasals may also be particularly challenging.

Turning to Table 2, it can be seen that the initials in Mandarin comprise a set of 24 consonants (Lee and Zee, 2003). There is also an argument as to the number of consonants in Mandarin. Duanmu (2007) suggested that there are only 19 consonants with three palatals and some syllabic consonants. Others (e.g. Eme & Odinye, 2008; Cheng, 2011) suggest there are 21 consonants in total. The education system in mainland China teaches 23 initials to students

(Hu, 2015). However, the most complete work so far is the one done by Lee and Zee (2003) and Table 2 is based on this.

*Table 2* Mandarin initials (consonants) written in Pinyin with IPA form in brackets. Initials marked in red are used in the current experiments.

	Unaspirated	Aspirated	Nasal	Voiceless fricative	Voiced fricative	Approximant
Bilabial	<b>b</b> [p]	p[p <sup>h</sup> ]	<b>m</b> [m]	f[f]		<b>w</b> [w]
Alveolar	<b>d</b> [t]	t[t <sup>h</sup> ]	<b>n</b> [n]		l[l]	
Velar	<b>g</b> [k]	<b>k</b> [k <sup>h</sup> ]	ng[ŋ]*	<b>h</b> [x]		
Palatal	<b>j</b> [t <sup>ɕ</sup> ]	<b>q</b> [t <sup>ɕʰ</sup> ]		<b>x</b> [ç]		<b>y</b> [j]
Dental sibilant	<b>z</b> [ts]	<b>c</b> [ts <sup>h</sup> ]		s[s]		
Retroflex	<b>zh</b> [tʂ]	<b>ch</b> [tʂ <sup>h</sup> ]		<b>sh</b> [ʂ]	r[ʀ]	

It should be noted that although the consonants w[w] and y[j] may be considered as consonantal allophones of the Mandarin vowels u[u] and [y]. They can appear at the beginning of a syllable followed by the corresponding vowel (glide), acting as an epenthesis rather than a separate phoneme. It may be perceived as both [wu] and [u] to a listener. These two allophones can also appear in the middle of a syllable. For example, 鳖 (/biē/, tortoise) can be written as [bie] or [bye]. Again, both versions of pronunciation may be perceived by a listener (Cao, 2016). Again initials marked in red in the table are those used in the stimuli for the current experiments. Palatals (**q**[t<sup>ɕʰ</sup>], **x**[ç]) and retroflexes (**zh**[tʂ], **ch**[tʂ<sup>h</sup>] & **sh**[ʂ]) are expected to be particularly difficult for English speakers since they do not exist in English, and this has been confirmed experimentally study (Ni & Wang, 1992).

Turning to tones: there are four basic tones, as well as a short and weak neutral tone. As suggested above, these tones are associated with the entire syllable at the supra-segmental level. The most frequently used system in describing Chinese tones is scale of five pitch levels developed by Chao (1930) (quoted from Sun, 2006). Basic information for each tone is summarized in Table 3 along with the standard diacritics used to represent tone in pinyin. Their pitch value change is also presented in Figure 1. Note that although phonologically the tone is associated with the entire syllable, in pinyin, diacritics are written upon the final (vowel); if the vowel is compound, then it is placed upon the final vowel (e.g. /huān/ rather than /hūan/). The current study used the first four tones. The 5th “neutral” tone is sometimes referred to as the lack of tone, and is used where the syllable is pronounced as a weak syllable. This can only happen on the last syllable in a word in Mandarin words with at least two syllables, or at the end of a sentence. In the current study, only one syllable words are used so that the 5<sup>th</sup> tone doesn’t occur<sup>3</sup>.

Table 3 Mandarin Chinese tones

<b>Tone Number</b>	<b>Marking in Pinyin</b>	<b>Pitch value</b>	<b>Description</b>
1 <sup>st</sup>	/ā/	55	High level tone (flat tone)
2 <sup>nd</sup>	/á/	35	High rising tone
3 <sup>rd</sup>	/ǎ/	214	Falling rising tone (dipping tone)
4 <sup>th</sup>	/à/	51	Falling tone
5 <sup>th</sup>	/a/		Neutral tone

<sup>3</sup> It should be noted that the condition of the 5th tone is relatively complicated in Mandarin and differs in different dialects. In some cases, using the neutral tone does not change the meaning of the word, thus it is optional (e.g. 因为 /yīn wéi/ or /yīn wei/, because). However there are also words where using the neutral tone does result in different meaning (e.g. 地道 /dì dào/, tunnel; /dì dao/, typical). In some cases there is evidence of language change since originally neutral tone was obligatory to be applied but new generation Chinese speakers do not apply the change as it does not change the meaning of the word (e.g. 儿子 /ér zi/ or /ér zǐ/, son) (Liu, 2013).

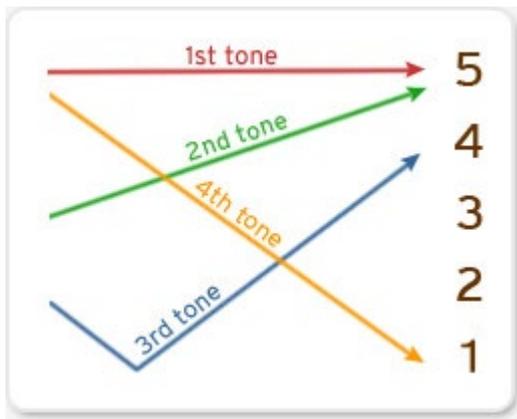


Figure 1 Mandarin Chinese Lexical Tones (ChinesePod, 2019).

Since the key question of this thesis is the learning of Mandarin tones by native English speakers, I devote the next section to understanding how native English speakers might perceive Mandarin tones.

### 1.3 Perception of Mandarin Tones by Native English speakers

Recall from Section 1.1 that both the SLM (Flege 1995) and the PAM (Best, 1994, 1995) suggest that difficulty of learning L2 phonetics is largely dependent on the degree of similarity between the L2 category and previous experience (L1 or other languages learned). They suggest that when participants discriminate non-native contrasts, they tend to assimilate these items to L1 categories. Although lexical tone does not exist in English, So and Best (2008) examined whether English speakers can assimilate Mandarin tones into familiar *intonational* categories. They hypothesized that on the basis of phonetic similarity with pitch contours of English intonation patterns, participants would perceive Tone 1 as *Flat Pitch*, Tone 2 as *Question*, Tone 3 as *Uncertainty*, and Tone 4 as *Statement*. The results supported 3 out of the four hypotheses. Only perceiving tone 3 as uncertainty was not supported as tone 2 and tone 3 were equally assimilated to uncertainty. However, the assimilation of tones was not mutually exclusive. For example, tone 1 was also frequently interpreted as “Statement” as well. Later,

So and Best (2011) built on this with a paradigm using sentence context and again explored whether English and French speakers would assimilate Mandarin tones into their intonational categories from their L1's. They found that while English speakers assimilate tone 3 and tone 4 into the same category, French speakers were able to spot the difference and categorise them differently. They hypothesised that this was due to the fact that French, but not English, is a syllable timed language like Mandarin. Thus French speakers are better at perceiving tones in the sentence context. The authors concluded that the PAM can be extended to suprasegmental phonology and that lexical tones are assimilated to categories of listeners' native phonetic system, but that rhythmic properties could also affect the perception of tone. This may make learning Mandarin particularly hard for English listeners. So and Best (2014) further built on their 2011 results by including a discrimination task alongside categorisation. Again, they found that tones were primarily categorised according to similarity to phonetic categories in native language and that also T3-T4 were easier for French than English listeners. In the discrimination task, they found that again French listeners outperformed English speakers. They also found that for both groups, performance was the lowest on T1-T4 and T2-T3 contrasts, with below 62% for English speakers and below 67% for French speakers.

Hao (2014) investigated the application of the SLM to Mandarin tone learning. Instead of using an assimilation task, he asked participants with no, little (1.5 year on average) or extensive (5.6 year on average) Mandarin experience to judge the English-likeness of the four Mandarin tones carried by 36 Mandarin monosyllables. The participants also took an identification tasks where they need to identify the tones of 36 Mandarin monosyllables repeated three times. The results of the identification tasks revealed an advantage for experienced learners on T1, T3, and T4, but not on T2, which implied that there was no learning effect for T2. This was also the tone which was judged to be the most similar to English by the

experienced learners. A limitation of the study was that ‘learning’ was represented as the difference between inexperienced participants and experienced participants, which can only roughly reflect individual learning outcomes. In addition, these researchers did not examine whether these English-likeness rating predicted participant’ identification performance (e.g. using correlation or regression). In addition, participants’ performance did not always straightforwardly fit with an account in which dissimilar tones are easier. For example, inexperienced learners rated tone 3 the least English-like tone, but their performance on T3 was actually the worst, suggesting that this tone maybe particularly difficult to learn, even though it is very different from English. Also, although numerically the experienced participants rated all four tones less English-like than no-experience group there was no statistical analysis of the difference between groups. Nevertheless, this study on the SLM, along with those conducted by researchers investigating the PAM, provides some evidence that as for the L2 segmental phonology, the perception of lexical tones by English speakers is influenced by their L1 knowledge.

From the evidence above, it can be seen that the speech perception models (PAM & SLM) which were originally developed for understanding the development of L2 learning can be extended to the learning of lexical tones. The current study focuses on native English speakers, who don’t have any prior knowledge of any tonal language. In light of the discussion above, I do not expect them to be completely unable to discriminate Mandarin tones: as suggested in the PAM model (So & Best, 2011), at least some aspects of their suprasegmental knowledge in the L1 will be able to assist the discrimination process. On the other hand, we have also seen that some aspects of English such as rhythmic properties make tone learning harder compared with speakers with other non-tonal language as their L1. However, it should be noted that the purpose of the current thesis is *not* to test the different predictions of PAM

and SLM (in which case I would focus more on the ability to differentiate, and potentially train *different* tonal contrasts; I will return to this point later in the general discussion (Section 5)). Instead, this thesis focuses on whether the ability to identify and discriminate Mandarin lexical tones can be trained in native English speakers more generally. In particular, there is a practical applied goal in understanding the extent to which this ability can be developed using phonetic training, and which type of training input might lead to better learning results. Specifically, I explore whether high variability phonetic training materials are helpful, and whether this depends on the individual differences of the learner.

#### 1.4 *High Variability Phonetic Training*

##### 1.4.1 *Phonetic Training of 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. The majority of this literature has considered segmental phonology and this is reviewed in the current section, turning to training of lexical tones in the following section.

An early study by Strange and Dittman (1984) attempted to train Japanese speakers on the English /r/- /l/ distinction, a phoneme contrast that does not exist in Japanese. Participants were trained on stimuli from a synthetic *rock-lock* continuum. The key result was that although performance increased both for trained and novel synthetic items, participants failed to show any improvement for naturally produced minimal pair items. Later research suggested that a key factor which prevented generalisation 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 only one single phonetic context and one single (synthesised) speaker. Logan, Lively and Pisoni (1991) also trained Japanese learners on the English /r/-/l/ contrast, but included multiple natural exemplars spoken by six speakers, with the target speech sounds appearing in a range of different phonetic contexts. In contrast to

Strange and Dittman, they found that participants successfully generalised both to new speakers and new words at test. This was the first study to indicate the importance of variability within the training materials. A follow up study by Lively, Logan and Pisoni (1993) provided further evidence for this by contrasting a condition with *high variability* input to one with *low variability* input in which the stimuli were spoken by a single speaker (although still exemplified in multiple phonetic contexts). Participants in the low variability group improved during the training sessions but failed to generalise this learning to a new speaker.

Following Lively et al. (1993), HVPT has become standard in L2 phonetic training. 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/ contrast by native Greek speakers (Giannakopoulou, Uther & Ylinen, 2013; Lengeris & Hazan, 2010), and learning of the English /w/-/v/ distinction by native German speakers (Iverson, et al., 2008).

There is also some evidence that this type of perceptual training benefits production in addition to perception. Bradlow, Akahane-Yamada, Pisoni and Tohkura (1999) 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 speakers following HVPT was also reported by Lambacher, Martens, Kakehi, Marasinghe, and Molholt (2005). However, the evidence here is mixed: A recent study by Alshangiti and 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.

Although the studies reviewed above all used HVPT, only the original work by Logan and colleagues directly contrasted the use of high and low variability materials. It is notable

that these seminal experiments used small samples (the tests of generalisation were administered to only three of the participants in Logan et al., 1991). Since then, few studies have explicitly contrasted high and low variability training. One such study was Sadakata and McQueen (2013), who 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). Both types of training led to increases in generalisation to untrained fricatives and speakers. However, in an identification task, the improvement was greater for participants receiving high variability training than those receiving low variability training. Similar results were reported by Wong (2012, 2014) who trained native Cantonese speakers with the English /e/ - /æ/ contrast. Both low variability (1 speaker) and high variability (6 speakers) training lead to increased performance from pre- to post- test, but the improvement was greater for the high variability group. This was found in tests of generalisation to new speakers and new items, and from perception to production. In contrast, a recent phonetic training study did not find the same benefit. 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/-/ɪ/ contrast. This study did *not* show a benefit for high variability compared to low variability training in either age group, even for generalisation items. However, for adult participants, it is unclear the extent to which this was due to ceiling effects. Two other studies which specifically manipulated variability during learning of non-native phonetic categories did so in the context of training lexical tone: Perrachione, Lee, Ha and Wong, (2011) and Sadakata and McQueen (2014). These studies are discussed in more detail in the following section.

Although there is a relatively small evidence base regarding a benefit of high over low variability phonetic training for non-native phoneme categories, there is further evidence for

this benefit in related areas of speech and language learning, such as accent categorisation and adaptation (Bradlow & Bent, 2008; Clopper & Pisoni, 2004). There is also a series of studies suggesting L2 vocabulary learning may benefit from high variability. Barcroft and Sommers (2005) trained native English speakers with 24 Spanish words. They found that participants who were exposed to more speakers or more voice types during training performed better than those only hearing one speaker or one voice type in a later assessment of word recall in both production and comprehension. A later follow-up study established that this variability benefit also held when the learning was performed in noise condition, with the advantage of multiple speakers increasing systematically as signal/noise ratio decreased, highlighting the benefit of hearing multiple speakers during training (Sommers & Bancroft, 2011). Sinkeviciute, Brown, Brekelmans, & Wonnacott (2019) replicated this benefit of multiple talker input for adult learners, although not in children (7 and 11 year old). Further studies by Barcroft & Sommers (Sommers and Barcroft (2007), Barcroft and Sommers (2014)) established that other variability manipulations, such as speaking rate, could also boost vocabulary learning, although which were effective could differ for speakers of different native languages.

In all of these areas, benefits of HVPT are generally seen in tasks of generalisation, suggesting that exposure to variation across speakers and/or items boosts the ability to generalise across these dimensions. This intuitively sensible result is in line with the predictions of computational models in which irrelevant contextual/speaker identity cues compete with phonetically relevant cues, so that dissociation of these irrelevant cues is the key mechanism which underpins generalisation (Apfelbaum & McMurray, 2011; Ramscar & Baayen, 2013; Ramscar, Yarlett, Dye, Denny & Thorpe, 2010).

#### 1.4.2 *Phonetic Training of L2 Lexical Tones*

Each of the phonetic training studies discussed above involved training a *segmental* contrast (consonantal or vocalic). Compared with such literature, studies on training

participants with lexical tones are quite limited. The first such study 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 materials were all real monosyllabic Mandarin words that varied in the consonants, vowels and syllable structures. Participants were native English speakers who were university students studying Mandarin with course learning experience from 4 to 10 months. During training participants heard a syllable whilst viewing two of the four 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 trial by trial feedback was provided. At test, participants heard a word and identified the tone out of a choice of all four diacritics. Generalisation items were included to test generalisation of untrained words and a new speaker. Participants showed significant improvement in the accuracy of tone identification after eight sessions of high variability training conducted over two weeks, and this generalised to both new words and a new speaker. In a follow up study, Wang, Jongman and Sereno (2003) used a similar high variability training paradigm to Wang et al. (1999) 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 - see section 1.2 above) before and after training. They found improvements in production, although these were mainly seen in pitch contour rather than pitch height. A further study explored the how individual differences might contribute to phonetic training with Mandarin tones. Wong and Perrachione (2007) trained native English speakers with no previous experience of tone languages, using an artificial lexicon of English pseudowords with superimposed Mandarin tones 1, 2 and 4. They employed relatively low variability training compared with previous studies as participants only heard one speaker during the training session. Training was set up such that participants learned to associate 18 pictures with 18 words. Although this set included some minimal pairs which were only

differentiated by tone, in contrast to other phonetic training studies, they did not train participants to distinguish these minimal pairs. Instead, these 18 words were separated into 6 groups such that within each group all words *differed* both in tone and segmental phonology. Training was carried out group by group: participants heard each word-picture combination repeated 4 times after which they heard a word and select the correct picture out of three from the same group. At the end of each training session, there was an identification task where participants had to match a target word from a choice of all 18 pictures- i.e. including some minimal pairs. Participants continued to receive one training session per day until they reached 95% accuracy on the identification task for two consecutive sessions (classified as “successful learners”) or show improvement smaller than 5% for four consecutive sessions (classified as “less-successful learners”). Note that 95% improvement would require them to be able to differentiate the minimal pairs using tone. However this study did not measure generalisation since the same talkers were used in the training and in the identification test. However, they did find while all participants showed significant improvement in identifying the words, their learning outcomes correlated with both previous musical experience *and* their ability to identify Mandarin tones on an additional tone identification task. This latter was a *Pitch Contour Perception Test* (PCPT) (a version of which task is used in the studies reported in the current thesis) in which participants heard a vowel produced with either Mandarin tone 1, 2 or 4 whilst viewing pictures of standard diacritics associated with these tones (→, ↗ & ↘), and were asked to select the arrow that corresponded to the tone. Further research has investigated the neurological underpinnings of these findings. Wong, Perrachione and Parrish (2007) used the same training paradigm as Wong and Perrachione (2007) and again trained naïve English speakers. As in the previous study, at the end of training, participants were classified as successful learners or less-successful learners based on their performance in training. Although both groups demonstrated significant improvement during training, participants classified as

“successful” showed increased activation in the left posterior superior temporal region after training, an area linked to the functioning of linguistic knowledge. In contrast, participants classified as ‘less-successful’ showed increased activation in the right superior temporal region and right inferior frontal gyrus which are associated with non-linguistic pitch processing, and prefrontal and medial frontal areas which are associated with increased working memory and attentional efforts.

Chandrasekaran, Sampath & Wong (2010) trained native English speakers with English pseudowords accompanied by 4 Mandarin tones. The training used 6 English pseudo words with four tones and they divided these 24 words into four groups. Each group was minimally contrasted only by tone. Each word was repeated by two male and two female speakers and each of these was repeated four times. Similarly to Wong & Perrachione (2007), after each group was presented, a quiz was given on the words just learned and at the end of each training session the identification task was carried out for the whole set of 24 words. They also tested participants using the *Woodcock-Johnson Tests of Cognitive Abilities*, which includes measures of phonological awareness and auditory working memory (including measures of digit span and letter number sequencing discussed further in Section 4.2.1) as well as including a pitch identification test (a version of the *Pitch Contour Perception* test described above) and a pitch discrimination test where participants had to discriminate different tones imposed on the English /a/ vowel. In the analysis, they separated participants into “good learners” vs “bad learners” on the basis of their performance in the last (9<sup>th</sup>) training session. Good learners showed clear advantage over bad learners in the pitch identification task, but *not* in the pitch discrimination task and *not* in any of the tasks measuring cognitive abilities. A follow up study explored the neural underpinnings of these differences. Chandrasekaran, Kraus and Wong (2011) again trained naive native English speakers with English pseudowords accompanied by four Mandarin tones with multiple speakers used in training. Again words were associated with

objects and participants were trained to associate words with objects, which required them to use lexical tone. At the end of each training session, participants were asked to match the word they learned with the object. Once more, participants showed significant improvement and the extent of improvement was linked to the activation of the inferior colliculus, a primary midbrain structure involved in representing auditory communication signals. A further lexical tone training study also looked at the role of individual differences, focusing on the relationship with musical ability. Li and DeKeyser (2017) trained native English speakers, with no previous experience of any tonal language, on Mandarin tone words. Participants were taught some Mandarin vocabulary and then received focused practice using either production or perception tasks. Production tasks involved reading pinyin and perception tasks involved matching the word heard with correct pinyin and picture. After training, all participants were tested on both perception and production. The results showed that participants' performance was far worse when tested on the skills that had not been trained. Critically, the study also included a set of tests of musical tonal ability using the Pitch Change Test measuring pitch perception ability, the Perceptive Tonal Memory Test measuring tone differentiation ability (both from Wing Measures of Musical Talents, Wing, 1968) and a Productive Tonal Memory Test in which participants need to reproduce a list of tunes 2-7 notes long (Slevc & Miyake, 2006). An overall musical tonal ability score was computed which combined these score using exploratory factor analyses. This score was found to correlate with both overall tone word perception accuracy and overall production ratings, regardless of training condition, indicating a role for musical ability in learning lexical tone.

These studies suggested that, as with segmental phoneme contrasts, phonetic training may also facilitate the learning of tone contrasts. In addition, individual difference may contribute to such learning process. However, although some papers reported above employed HVPT in the sense of using multiple speakers, none of the studies reported so far directly

contrasted high and low variability training materials. In addition some of the studies focusing on individual differences did not examine whether the improvement in training can generalise to new stimuli. Recently, two studies (Perrachione, et al., 2011; Sadakata & McQueen, 2014) looked at these questions specifically. Further details of these studies are given in the next chapter (section 2.1.1) since they are the direct inspiration for Study 1 and Study 2. Perrachione et al. (2011) used a similar paradigm to that used by Wong and Perrachione (2007) where native English speakers with no exposure to tonal language were trained to associate tones combined with English pseudo words with objects. They directly compared *high* (four speakers) and *low* (one speaker) *variability input* training results. In addition, they used the *Pitch Contour Perception task* (described above) as a separate measure of learner aptitude, specifically they used this measure to divide learners into “high” and “low” aptitude groups based on cut off values from Wong & Perrachione (2007). The key finding is that they did not find an overall benefit of high variability, however, they did find an interaction between participant aptitude as measured by the PCPT and variability condition: participants with high aptitude benefited more from high variability (HV) training while low aptitude participants only benefited from low variability (LV) training. Sadakata & McQueen (2014) found evidence for a similar interaction in a rather different training study with native Dutch speakers, again with no experience of learning a tone language. Their training paradigm required participants to associate Mandarin lexical tone contrasts with numbers (e.g. Tone 3 – Tone 1 labelled as ‘1’). Three variability conditions (high, medium and low) were used with both speaker and item variability manipulated. Their measure of individual aptitude was a categorisation task using stimuli from a six step Tone 2 to Tone 3 continuum where participants were asked to identify if the sound they heard was more like Tone 2 or Tone 3. Categorisation slopes for the participants were used as a measure of their ability to discriminate this contrast and on the basis of this test participants were again divided into groups of high and low aptitude learners. The

important finding was that there was no overall benefit of being in a higher variability condition, however there was an interaction between input variability and learner aptitude, with high aptitude participants benefiting more from higher variability materials and low aptitude participants benefiting from low variability materials.

The results of these two 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.

### ***1.5 Overview of the current thesis***

From the literature above, it can be seen that researchers have only relatively recently tried to apply the HVPT methods on Mandarin learning and the amount of research is relatively limited. The current thesis continues this line of research, further investigating the role of high versus low variability in training materials and the role of individual differences in learning tonal languages.

Chapter 2 reports a training study (Study1) with English native speakers which directly compares training with a single speaker (low variability - LV) versus multi-speaker (high variability - HV) materials. It builds directly on the studies by Perrachione et al. (2011) and Sadakata and McQueen, (2014), but aims to extend their finding to an experiment using training materials involving all four Mandarin tones embedded in real Mandarin words. In this study, participants' individual aptitudes were measured using tests of tone identification and categorical discrimination based on those used in the two previous studies. Significant improvement during training and successful generalisation to new speakers and new items was found for both LV and HV training groups, on both production and perception tasks. However, in contrast to the previous studies, while participants with higher aptitude (at least as measured in one of the aptitude tasks, *Pitch Contour Perception Test*) tended to performed better in

baseline measures of tone discrimination, there was no evidence that performance on these aptitude tasks predicted the extent of learning from the training materials and, crucially, there was no interaction between variability condition and individual aptitude. There was also no evidence of an overall HV benefit, contradictory towards what has been reported in previous literature comparing HV and LV input (Logan & Pisoni 1993).

Chapter 3 reports a study (Study 2) which essentially adds an extra condition to the experiment reported in Chapter 2. One concern with the high variability condition used in Study 1 was that it was potentially more difficult for learners due to speakers varying trial by trial. The new condition controls for trial-by-trial consistency during training in a new version of the high variability condition where the stimuli from each speaker were presented in blocks (high variability blocking (HVB)). In training, the performance of this new HVB group was higher than the HV group and was now at a similar level to the LV group. However, aptitude (measured with *Pitch Contour Perception Test*) was still not predictive of learning and there was still no interaction between variability and individual aptitude and no overall advantage when training participants with lexical tones using HV materials, even controlling for trial-by-trial consistency.

In addition, in Chapter 3, a new type of statistical analysis was conducted in order to further investigate the null results seen in Study 1 and 2. A limitation of the *p*-value, the inferential statistic used up to this point in the thesis, is that it does not allow us to differentiate the finding that there is no evidence for an effect, from the finding that there is evidence *for* a NULL result. Thus, an additional set of analyses using Bayes Factors (BFs) was run, making it possible to quantify evidence for the null. BFs showed that there was substantial evidence for the null for the hypothesis that high variability promotes generalisation. However for the hypothesis of an interaction between variability and aptitude the evidence was ambiguous.

Further analysis suggested that the paradigm wasn't sufficiently sensitive to test this hypothesis with these individual difference measures, at least not without a very large sample.

One of the limitations from studies reported in Chapter 2 and 3 was that although one of the aptitude measures (*Pitch Contour Perception Test* adapted from Perrachione et al. (2011)) did correlate with baseline performance measures in the experiment, there was no evidence suggesting it predicted the extent to which participants could benefit from training (i.e. pre- to post improvement). Chapter 4 (Study 3) further probes the types of measures of "individual aptitude" which may predict learning in HVPT. In addition to the *Pitch Contour Perception Test* used in the previous study, measures of working memory, attention and musical ability were included. Study 3 only used multi-speaker training paradigm as the strongest correlation was expected with measures of individual differences (*ID measures*) and this specific training condition. This study also explored whether the role of different *ID measures* in predicting learning was different for naïve learners and participants who were already students of Mandarin in the university. The results suggested interesting differences between the two groups: learning for the participants who had previous knowledge of Mandarin was primarily predicted by attention measures, while this was not the case for naïve participants, where learning was mainly predicted by musical ability measures and to some extent working memory (*Digit Span*) measures. Again, BF analysis was performed, allowing us to quantify where there was evidence for the null.

Finally, chapter 5 summarises and discusses the main findings of the current thesis. It focuses on both the lack of advantage for high over low variability training seen in the first two studies, and discusses what role exactly individual differences might have when learning Mandarin tone contrasts. Theoretical and methodical implications are discussed, followed by

possible future research directions, including an experiment which would return to compare HV and LV input, as in Study 1 & 2, but using the types of *ID measures* evaluated in Study 3.

## 2. Study 1

### 2.1 Introduction

As discussed in section 1.4.2, although HV materials have been suggested to be effective in training Mandarin tones (Wang et al., 1999; Wang et al., 2003), relatively few studies have directly compared LV and HV training. This is true both within the broader phonetics training literature and for tones specifically. Where HV and LV materials have been compared in the context of tones, individual difference seems to play a crucial role, with two studies (Perrachione et al., 2011; Sadakata & McQueen, 2014) finding an interaction between “individual aptitude” and the benefit of variability in training. This chapter presents a study which is a conceptual replication and extension of these two previous studies. First, I further describe the details of the studies, before turning to discuss the design decisions in the current study.

#### 2.1.1 *The studies by Perrachione et al., 2011 and Sadakata & McQueen, 2014*

Perrachione et al. (2011) 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 LV (one speaker) or HV (four speaker) input. During the training, participants matched the sound they heard with one of three pictures of concrete objects presented, where the three words associated with these pictures were minimal trios that differed only in tone. Participants were tested on their ability to generalise their learning to new speakers. Importantly, they determined participants’ individual aptitude, i.e. the baseline ability to perceive the tone contrasts prior to training using a *Pitch Contour Perception Test*. As described in the general introduction, this test had been used in previous studies (Wong & Perrachione, 2007; Wong, Perrachione and Parrish, 2007) and asks participants to match a tone produced across a single vowel to a picture of a diacritic

(→, ↗ & ↘). Based on performance in this task, the researchers grouped participants into high and low aptitude groups. The results showed that whilst the low variability group outperformed the high variability group during training (presumably due to accommodation to a repeated speaker throughout the task), there were no difference between the high and low variability groups during test. Critically, however, there was an interaction between an individuals' aptitude categorisation and input variability: only participants with high aptitude benefitted from high variability training, while those with low aptitude actually benefitted more from low variability training. It is important to note that this interaction was seen in a generalisation task with a similar form as training, but with a novel speaker. Somewhat confusingly, Perrachione et al (2011), do *not* refer to this task as a generalisation task and instead reported a generalisation measure which was a ratio of performance on this test with novel speakers to performance in the last training session (test-performance/training-performance). Note that this ratio would increase not only if one group of participants were better at test, but also if they were *worse* in training. Using this measure, they found a benefit of high variability training. However on inspection of the means, it seems that this relationship is driven by the *poorer* performance in training in the high variability condition, rather than by *better* performance in the test with novel speakers. Therefore, I suggest that this ratio measure should not be considered as providing evidence for an overall benefit of HV training on generalisation. The fact that performance with novel speakers was not better with HV materials is surprising in the context of previous literature in phonetic training discussed in section 1.4.1 (e.g. Lively et al., 1993) and computational models (Apfelbaum & McMurray, 2011; Ramscar & Baayen, 2013) suggest that exposure to multiple speakers should be specifically beneficial in generalisation since it should allow learners to better dissociate the tones from the particular speakers used in training. The interaction found by Perrachione et al. (2011) seem to suggest that only the high aptitude learners can take advantage of this benefit.

Sadakata and McQueen (2014), found a similar result using different training and testing materials. They trained native Dutch speakers (with no prior knowledge of Mandarin or any other tonal language) using naturally produced bi-syllabic 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 (e.g. for one participant Tone 2-Tone 1 was labelled “1”, Tone 3-Tone 1 was labelled “2”). During the training task, participants identified 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 1). 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. Specifically, the number of different vowels used in the bi-syllabic sequences was manipulated: the low variability group encountered only one vowel (.e.g. pasa, casa, lasa, etc.) whereas the medium and high variability groups encountered four different vowels (pasa, pesa, pisa, pusa; casa, cesa, cisa, cusa; lasa, lesa, lisa, lusa etc.). Participants were tested on the trained items (i.e. using trained speakers and trained items). Generalisation was also examined in a number of ways by looking at (1) trained items spoken by an untrained speaker; (2) pseudowords containing untrained vowels (3) pseudowords in which the order of tones in the bi-syllables were reversed (i.e. a novel position), and (4) items where the tone was embedded in a sentence context.

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 *Categorisation of Synthesized Tonal Continua* task using stimuli from a six step Tone 2 to Tone 3 continuum (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 categorisation 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 this grouping was entered as a factor in the analyses of tests of learning, along with the effect of training condition (high-medium-low) and the interaction between factors. For the test with trained speakers and items, there was no group level effect of variability condition, however there *was* an interaction between variability and aptitude similar to that reported by Perrachione et al. (2011): Participants with high aptitude benefitted more from high variability training, while those with lower aptitude benefitted more from low variability training. Note that however, here the interaction was found with trained speakers where as Perrachione et al. saw this interaction with untrained speakers. For the generalisation tests, participants showed above chance performance in all but the new position condition, demonstrating an ability to generalise their learning of tone across different dimensions. However, they did *not* demonstrate an overall benefit of higher variability in any of the transfer tests, nor did variability interaction with aptitude (i.e. unlike in Perrachione et al. (2011), where they found the interaction in generalisation). Note that the overall lack of a high variability benefit is again surprising, particularly for test items with untrained speakers and novel items, since the manipulations in training should specifically work to increase generalisation along these dimensions.

The fact that neither of the tone training studies found an overall benefit of high over low variability in tone generalisation is surprising in light of the phonetic literature and the predictions of the computational model (Apfelbaum & McMurray, 2011; Ramscar & Baayen, 2013) mentioned in section 1.4.1. Moreover, as the previous authors point out, if it is actually the case that learning from multiple voices is more or less effective for different groups of

learners, this has important implications for the design of L2 training tools. For this to be the case, it is important to establish the generalisability of the findings to different contexts and materials, particularly those which are relevant in an L2 learning context. Presumably, what Mandarin L2 learners are most interested in is developing their ability to use tones when mapping a word's phonological form to its meaning (and vice versa). In this light, the paradigm used by Sadakata and McQueen (2014) lacks ecological validity in looking only at mapping to abstract tone categories. On the other hand, Perrachione et al. (2011) do train form-meaning mappings, yet, unlike Sadakata and McQueen (2014) they use English pseudo-word stimuli, which has the consequence that learners do not simultaneously have to deal with non-native segments and tones, as in a real world L2 Mandarin learning situation. Furthermore, although there is limited data on the differences between words and non-words in production, it has been noted that non-words may have different properties from real words even within the same language (Scarborough, 2012) and may be more clearly articulated (Hay, Drager & Thomas, 2013; Maxwell, Baker, Bundgaard-Nielsen & Fletcher, 2015). Thus, using non-words might make stimuli slightly easier to learn than if real words were used.

### *2.1.2 The current study*

The current training study addressed these issues in a partial replication of the previous work: it used stimuli produced by native Mandarin speakers which were real words in that language. This design choice followed earlier studies such as Perrachione and Wong (2007) and Wang et al. (1999) using a paradigm in which participants were trained to identify word meaning on the basis of tone. However, in contrast to those studies, the current design trained the contrasts between all four tones (six tone contrasts) rather than just three (on the assumption that learners are interested in learning the complete set of contrasts within a particular language). It should be noted that these design choices potentially increased the difficulty of the training materials compared to previous work. A key question was whether these choices

would impact the interaction between learner aptitude and the benefits of more variable training materials.

Following Perrachione et al. (2011), this study only varied variability along one dimension – speaker variability, keeping training items identical across conditions. It also included two training conditions: low variability (one speaker) and high variability (four speakers intermixed within each training session). Two perceptual tasks designed to tap individual aptitude were used. These were adapted from those used in Perrachione et al. (2011) and Sadakata and McQueen (2014). However, while the previous studies grouped participants into one of two categories (high aptitude versus low aptitude) based on the aptitude score, in the current study they were used as continuous measures. This avoided assigning an arbitrary “cut off” for high versus 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 the current paper (logistic mixed effect models) made it possible to include continuous predictors and look at their interactions with other factors.

A further extension in the current study was that several new outcome measures were included to test learning and generalisation. First, most similar to the task used in Perrachione et al. (2011) was a picture identification task which was a version of the training task (2AFC picture identification) without feedback. Following Perrachione et al. (2011), it included untrained speakers, where benefits of speaker variability in training should be the most apparent. However, bearing in mind that Sadakata and McQueen (2014) actually found the key interaction with aptitude *only* in the test with trained stimuli, trained speaker test items were also included.

A second perceptual task was also included which did *not* involve knowing specific form-meaning mappings and thus had the benefit that it could be conducted both pre- and post-

test. This was a Three Interval Oddity task which required participants to pick the odd-one-out item after hearing three words spoken aloud, each by a different speaker. Two of the tokens were productions of the same word and the third differed only in the tone (e.g. *bā*, Tone 1; *bā*, Tone 1; *bà*, Tone 4). As all three tokens are physically different, it requires the listener to focus on the phonological level ignoring irrelevant acoustic differences. Furthermore, the use of three speakers forces the listener to ignore irrelevant speaker-specific differences, making it especially challenging (Strange & Shafer, 2008). This task used untrained speakers in every trial, so that every test-item required generalisation to new speakers. In addition, here it was possible to use both trained and untrained *items*. Note that even though the variability over items is matched across conditions, it is possible that varying speaker specific cues might also promote generalisation across this dimension. If this is the case, a high variability benefit may be stronger for untrained items than trained items.

Finally, this study tested production using a Picture Naming task at post-test, in which participants were required to name the pictures used in training in Mandarin. In addition, a Word Repetition task was conducted, which had the benefit that it could also be employed at pre-test, and that both trained and untrained items can be used. Although there is evidence HVPT can benefit the production of tones (Bradlow et al., 1999; Lambacher et al., 2005; Wang et al., 2003; Wong, 2014), there has been no direct examination of whether high variability training materials are more effective than low variability training materials for production. However, more generally in the L2 vocabulary learning literature, training with multiple speakers has been found to lead to better recall in a picture naming task (Barcroft & Sommers, 2005), suggesting that the HVPT advantage should extend to production measures.

In sum, the current experiment assessed whether individuals benefit from high over low variability perceptual training when learning novel L2 tone contrasts, and whether this interacts

with learner aptitude. Measures of aptitude were constructed based on previous studies, but the training paradigm employed real Mandarin stimuli embedded in a vocabulary learning task, which trained discrimination of all six Mandarin tone contrasts. Learning and generalisation were measured in multiple tests of both perception and production. In general, the current design increased ecological validity and likely also increased the difficulty of the learning task relative to previous work. It is possible that increasing difficulty could exacerbate differences between learners of different aptitudes, potentially increasing interactions with aptitude. On the other hand, it is also possible that the increased difficulty might make high variability input much harder for all participants, decreasing or removing the specific benefit of HVPT for high aptitude learners.

## **2.2 Method**

### *2.2.1 Participants*

Forty adults recruited from UCL Psychology Subject Pool participated in the experiment, twenty in each of the two conditions (LV, HV). Participant information is summarised in Table 4. There was no difference between these groups in age,  $t(38) = 2.74$ ,  $p = 0.37$ . 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 two were native speakers of English. Of these two, one participant (low variability condition) was a native bilingual of English and Hindi, one participant (high variability condition) was a native French speaker. Critically none had any prior experience of Mandarin Chinese or any other tonal language. On average, participants learned 2.1 ( $SD = 0.6$ ) languages and the average age for starting to learn the first L2 was 11.8 ( $SD = 0.9$ ).

Table 4. 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	Average
			Learned	Starting Age
Low Variability	26.15	19-53	2.7	13.8
High Variability	25.65	19-47	2.5	12.2

## 2.2.2 Stimuli

### 2.2.2.1 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 segments, differing only in tones (e.g. *māo*, Tone 1 [*cat*] versus *mào*, Tone 4 [*hat*]). The words were chosen to be picturable and to start with a wide range of segments (see Appendix A). In order to examine generalisation 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 5 summarises how speakers were assigned to each task.

In the LV condition only one speaker (Trained Speaker) 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 HV condition, four speakers were used in training. Only one of these speakers (Trained Speaker) was used in the Word Repetition test and for those trained speakers in the Picture Identification test (the same speaker across both tests). In both conditions, a further speaker (New Speaker) was assigned to the untrained speaker condition in the Picture Identification test. The assignment of speakers was rotated across participants, resulting in 5 counterbalanced versions of each condition (see Table 5). 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.

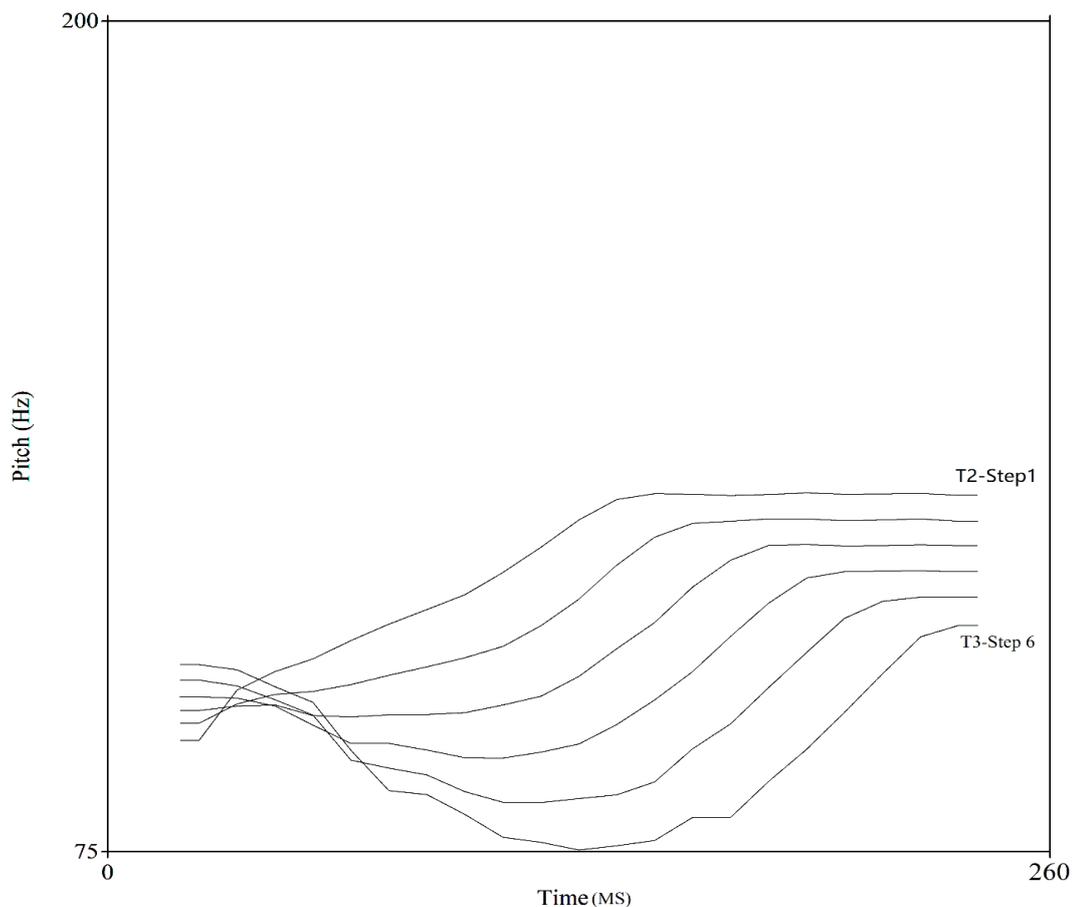
#### 2.2.2.2 *Stimuli used in the Aptitude Tests*

Pitch Contour Perception Test: Six Mandarin main 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 speaker set 2 described in the previous section) 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. An

English 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. *Figure 2* shows the contour patterns of these stimuli. 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.



*Figure 2* The pitch continuum from T2 to T3 used in Categorisation of Synthesized Tonal Continua

Table 5 Counterbalancing of voices for each task, training condition and version. LV = Low Variability; HV = High Variability; 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	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 Speaker	All	F1	F2	F3	M1	M2
New Speaker	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				

### 2.2.3 Procedure

The experiment involved three stages (see Figure 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. All sessions took place in a quiet, soundproof testing room in Chandler House, UCL.

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 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.



Figure 3 Tasks completed in each of the eight sessions (PCPT = Pitch Contour Perception Test; CSTC = Categorisation of Synthesized Tonal Continua).

### 2.2.3.1 *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 Categorisation of Synthesized Tonal Continua (described below) the test was also repeated at post-test and conducted analyses to identify whether performance on this task itself was improved as a result of training (see Section 2.3.2.1).

### 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 one 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). However, in line with their procedure, participants completed the task both *before* and *after* training and I conducted analyses to explore whether there was improvement from pre to post-test (see Section 2.3.2.2).

### 2.2.3.3 *Three Interval Oddity Test*

This task required participants 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 5). 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. 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 3).

### 2.2.3.4 *Word Repetition Test*

All 72 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 speaker; see Table 5). 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 2.3.5.1). This task was completed once in the pre-test and once in the post-test.

#### 2.2.3.5 *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 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.

#### 2.2.3.6 *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 4). 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 4).

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 in 288 trials in total per session. Participants were assigned to either LV or HV (with the assignment of speakers counterbalanced – see Table 5). Each session lasted for approximately 30 minutes.

In the LV condition, only *one* speaker was used. In the HV condition, *four* speakers were used. For these two conditions, all 288 trials were randomised so there was no fixed order of speaker in the HV condition. For each participant, each of their six training sessions was identical. After each block, the number of coins they had earned so far was displayed on the screen.

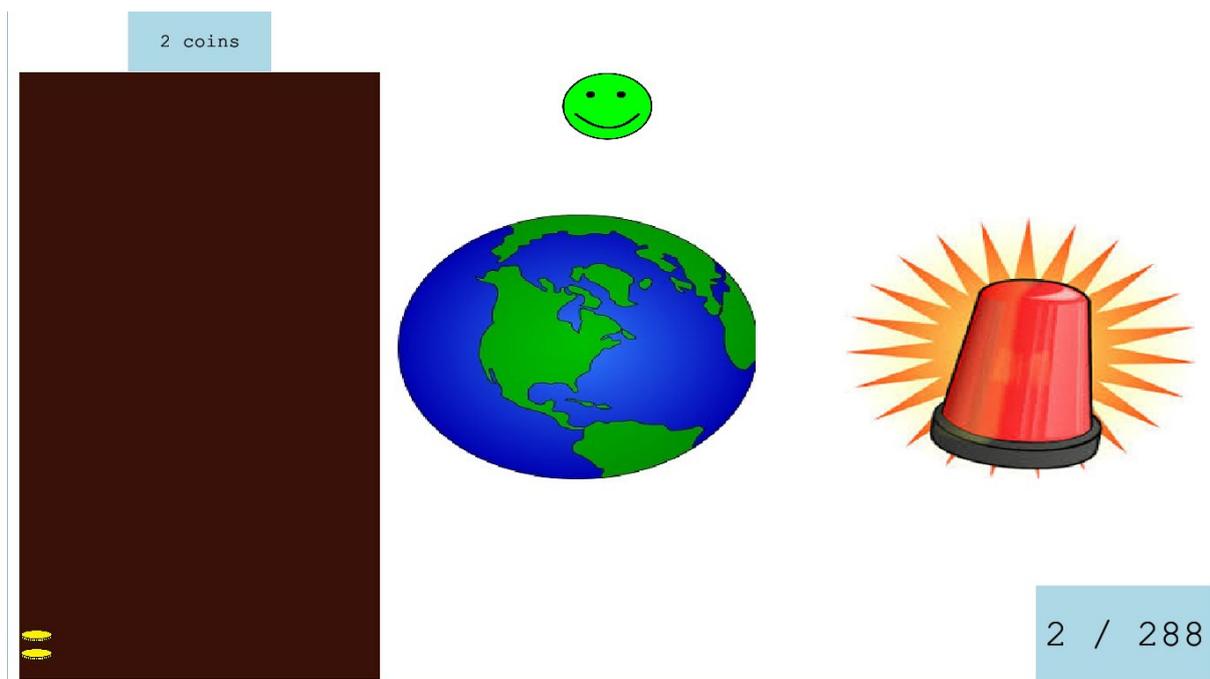


Figure 4 Screen shot from the training task. The stimuli heard is 'di', tone 4, [earth]. The foil picture on the right is 'di' tone 2, [siren].

### 2.2.3.7 *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 Speaker) and once by an untrained speaker (New Speaker), making 72 trials in total. Secondly, no feedback was given. This task was completed only in the post-test.

### 2.2.3.8 *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 2.3.5.1). This task was completed only in the post-test.

### 2.2.3.9 *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.

## 2.3 **Results**

### 2.3.1 *Statistical Approach*

Three different sets of frequentist analyses are reported. First, analyses were conducted on the data from the *Pitch Contour Perception Test* (Section 2.3.2.1) and *Categorisation of Synthesized Tonal Continua* (Section 2.3.2.2). The primary aim of these analyses was to ensure that the two groups did not differ at pre-test, however I also looked for possible differences at post-test. Second, separate analyses are reported for each of the tests administered pre- and

post- training (i.e. Word Repetition task (Section 2.3.5.2) and Three Interval Oddity task (Section 2.3.4.1)); the data collected during Training (Section 2.3.3) and the data from the two tasks administered only at post-test (i.e. the Picture Identification task (Section 2.3.4.2) and Picture Naming task (Section 2.3.5.3)). These analyses explored the effects of the 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 2.3.6). Specifically, the aim was 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 (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). Logistic mixed effect models allow binary data to be analysed with logistic models rather than as proportions, as recommended by Jaeger (2008). In all models, *variability-condition* was a factor with two levels (low, high) which was given a centered coding to ensure that other effects were evaluated as averaged for both levels of *variability-condition*. For the Three Interval Oddity task, a new factor, *trial-type* was also included. The purpose of this was to control for the fact that participants were likely to find some trial types easier than others due to the gender of the speakers producing the stimuli. The factor *trial-type* was coded into three levels (neutral, easy, hard - see Section 2.2.3.3) and included contrasts with neutral (“neutral versus easy” and “neutral versus hard”) 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 examination of 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 coded 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 *ID measures* in the models reported in Section 2.3.6), were centred (i) to reduce the effects of collinearity between main effects and interactions, and (ii) so that the 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). Experimentally manipulated variables and all of their interactions were automatically put 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 only its main effect and the interaction with *test-session* was used). Not all main effects and interaction coefficients were inspected within the models. Only those with the statistics which were necessary to look for accidental differences at pre-test, and those related to the current hypotheses were inspected and reported. The aim was to examine whether the training improved participants’ performance on both untrained items and untrained voices and whether such improvement was modulated by their individual aptitude. 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/j6s7w/?view\\_only=497e0e8ee7ff4e7387984690eafd4b5a](https://osf.io/j6s7w/?view_only=497e0e8ee7ff4e7387984690eafd4b5a)

## 2.3.2 Individual Aptitude Tasks

### 2.3.2.1 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 *variability-conditions* (LV, HV) and *test-session* (pre-test, post-test). There was no significant difference between conditions ( $\beta = -0.35$ ,  $SE = 0.26$ ,  $z = -1.38$ ,  $p = 0.17$ ) at pre-test on this measure. Participants showed a small yet significant improvement after training ( $\beta = 0.17$ ,  $SE = 0.06$ ,  $z = 2.68$ ,  $p < 0.01$ ), which can be seen in Figure 5.

Given that this measure is affected by training, only participants' scores at pre-test were used as the measure of individual differences in the analyses reported in Section 2.3.6.

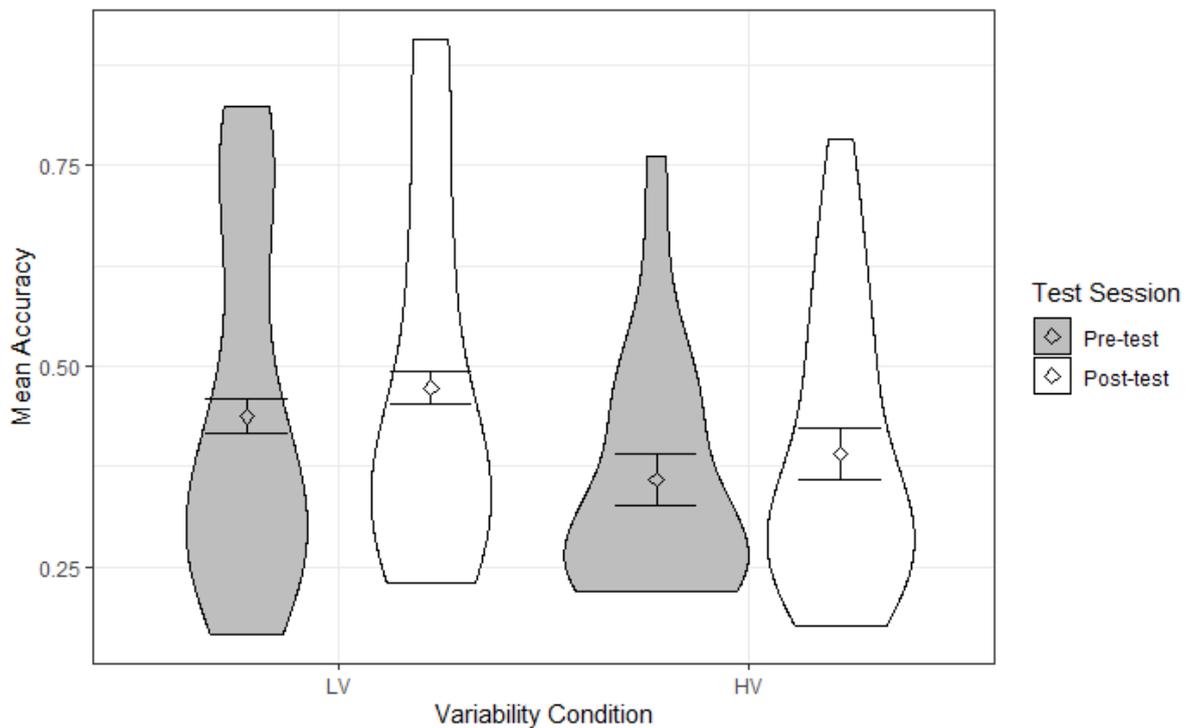


Figure 5 Mean proportion of correct for the LV (Low Variability) & HV (High Variability) groups in Pitch Contour Perception Test. Error bars represents the 95% confidence intervals.

### 2.3.2.2 *Categorisation of Synthesised Tonal Continua*

Following Sadakata and McQueen (2014), in order to quantify performance in this task, each subjects' categorization curve was fitted to a logistic function using the Logistic Curve Fit function in SPSS and a slope coefficient was calculated (Joanisse, Manis, Keating, & Seidenberg, 2000) which was taken to indicate the participants ability to categorize the two tones, with smaller (less steep) slopes indicating better performance (if participants categorise all trials correctly, the perfect slope coefficient they will get is 0.042). Sadakata and McQueen (2014), reported that they removed any participant with a slope measure greater than 1.2 from the analysis, suggesting that slopes above this threshold were considered to be poorly fit. However following this process with the current participants, the majority were above this threshold (33 out of 40). Given this, an alternative method was used. The slope coefficients for each participant was calculated using a logistic mixed effect model (Schultz, Llanos, & Francis, 2003) with the predicted variable being which of two tones the participants chose on each trial and the predictor being the tone step presented on each trial (varying from 1-6 where 1 is most like Tone 2, and 6 is most like Tone 3). Random intercepts and slopes for tone step were fit by participant and the individual slope coefficients for each participant were extracted from the by-participant random slopes fit in the model. As the random slopes represent adjustments to the fixed effect slope, more positive slopes represent sharper categorization responses, i.e. more sensitivity to differences in tone step, while more negative slopes represent flatter categorization responses, or in extreme cases reversed responses, and slopes close to 0 reflect responses close to the mean. These slopes could thus be taken to be indicators of individual differences. These slopes resembled the ones used by Sadakata and McQueen, as the same participants who failed their criteria also had very shallow slopes using the logistic regression method. The analyses conducted in the current study were done with the slope measure extracted from the logistic regression.

Again, a model was run to examine whether there was a difference between participants at pre-test and no significant difference was found ( $\beta = -0.43$ ,  $SE = 0.39$ ,  $t = 46.70$ ,  $p = 0.27$ ). There was no improvement of this measure after training ( $\beta = -0.01$ ,  $SE = 0.13$ ,  $t = 38$ ,  $p = 0.91$ ), which can be seen in Figure 6. It should be noted here that a negative slope indicates that participants categorise the stimuli in the wrong direction, i.e. they categorise more T2-related trials as T3, and vice versa.

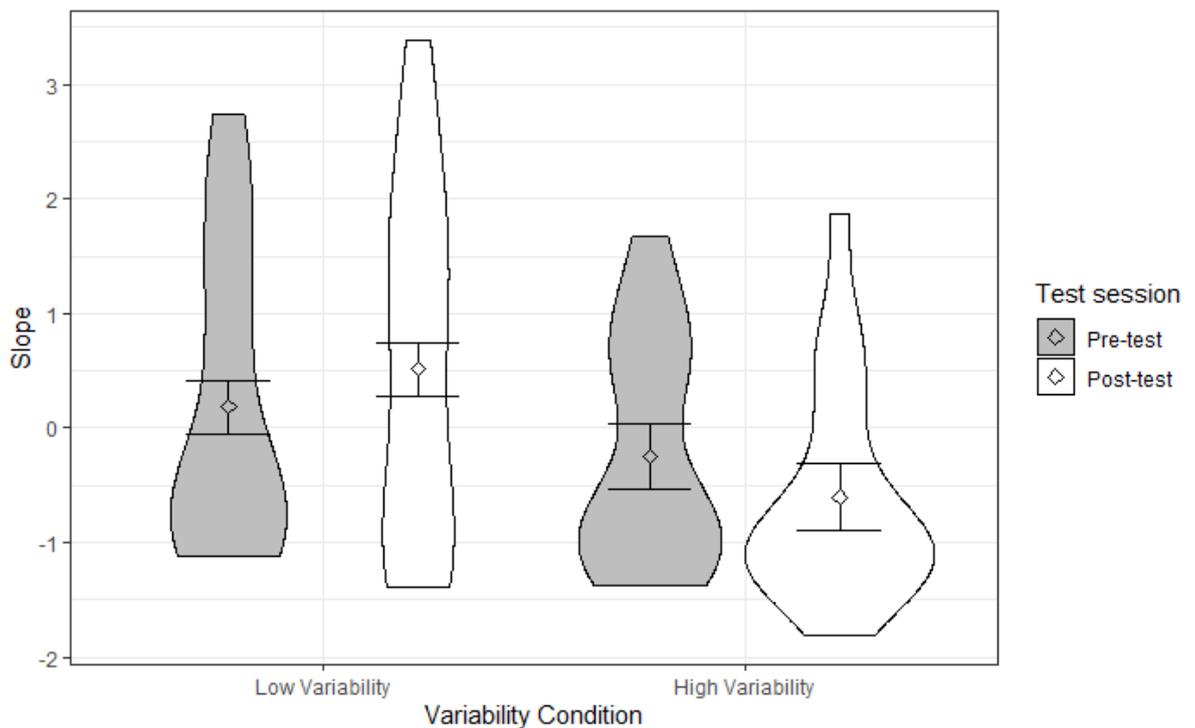


Figure 6 Slope measure for the LV (Low Variability) & HV (High Variability) groups in Pitch Contour Perception Test. Error bars represents the 95% confidence intervals.

In addition, a correlation analysis was run to look at the relationship between this ID measures and the one described in section 2.3.2.1 (the *Pitch Contour Perception test*). Results indicated a significant positive relationship between them:  $r(38) = .67$ ,  $p < .001$ .

### 2.3.3 Training

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 (LV & HV). The mean accuracy is displayed in Figure 7.

There was an effect of *training-session* ( $\beta = 0.49, SE = 0.04, z = 11.62, p < .001$ ): Participants' performance increased significantly over time, with additional training sessions. Overall, the LV group performed better than the HV group ( $\beta = -0.78, SE = 0.16, z = -4.95, p < .001$ ). There was a *training-session* x *variability-condition* interaction ( $\beta = -0.19, SE = 0.04, z = -4.54, p < .001$ ). From Figure 7 it can be seen that the LV and the HV group differed starting from the first session and this difference continued to increase throughout training.

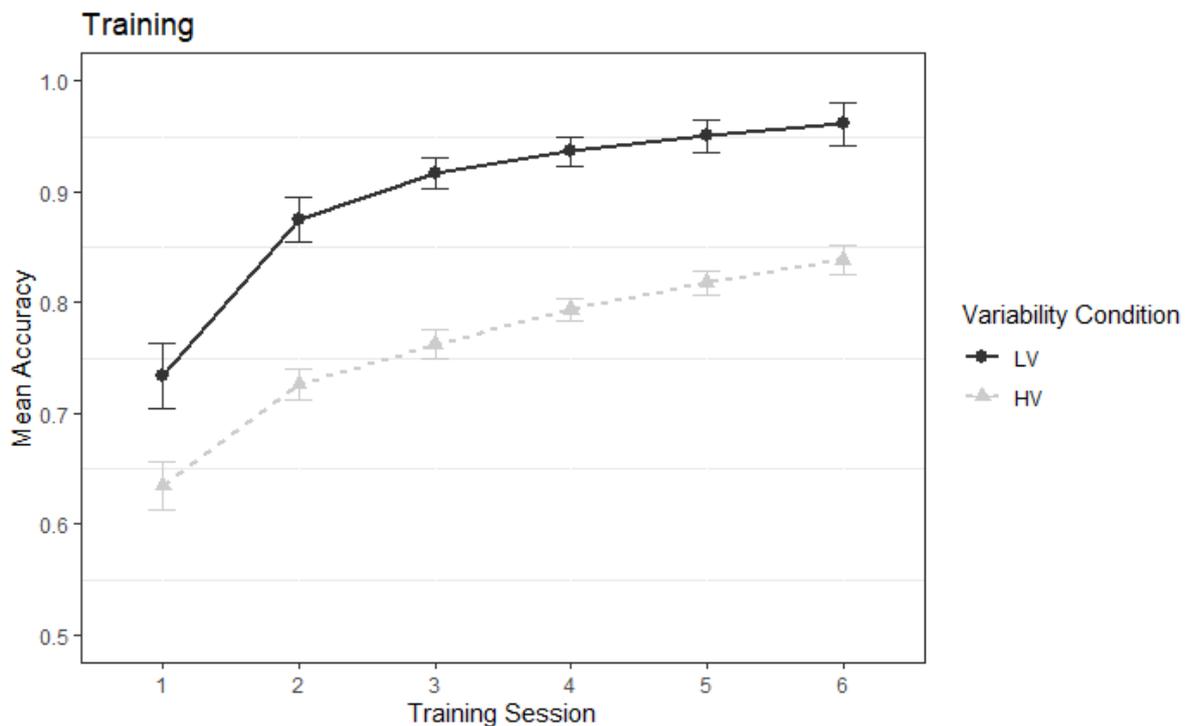


Figure 7 Mean proportion of correct in the Training task for the LV (Low Variability) & HV (High Variability) training groups in each session. Y-axis starts from chance level. Error bars show 95% confidence intervals.

### 2.3.4 Perceptual tests

#### 2.3.4.1 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, HV), *trial-type* (neutral versus easy, neutral versus hard) and *item-novelty* (trained item, untrained item). The mean accuracy is displayed in Figure 8.

At pre-test, there was no significant difference between the LV and the HV groups ( $\beta = 0.0003$ ,  $SE = 0.16$ ,  $z = 0.001$ ,  $p = .99$ ), suggesting that the groups started at a similar level. However, performance with the untrained items was significantly better than performance on the trained items at pre-test ( $\beta = 0.31$ ,  $SE = 0.08$ ,  $z = 4.01$ ,  $p < .001$ ), suggesting accidental differences between item sets. 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,  $\beta = 0.45$ ,  $SE = 0.10$ ,  $z = 4.70$ ,  $p < .001$ ) and “neutral” trials were easier than “hard” trials (where one of the foil speakers had the odd gender out) but the difference was not significant ( $\beta = -0.12$ ,  $SE = 0.09$ ,  $z = -1.23$ ,  $p = 0.22$ ).

Overall, participants’ performance increased significantly after training ( $M_{pre} = 0.58$ ,  $SD_{pre} = 0.19$ ,  $M_{post} = 0.65$ ,  $SD_{post} = 0.19$ ,  $\beta = 0.30$ ,  $SE = 0.06$ ,  $z = 5.26$ ,  $p < .001$ ). The interaction between *test-session* and *item-novelty* was marginally significant, with trained items improving more than untrained items ( $\beta = -0.21$ ,  $SE = 0.11$ ,  $z = -1.84$ ,  $p = 0.07$ ). However there was no difference between trained and untrained items at post-test ( $\beta = 0.12$ ,  $SE = 0.08$ ,  $z = 1.47$ ,  $p = 0.16$ ) and the accidental difference between trained and untrained items at pre-test (see above) makes this hard to interpret. Critically, there was no interaction between *test-session* and *variability-condition* ( $\beta = -0.01$ ,  $SE = 0.11$ ,  $z = -0.16$ ,  $p = .87$ ) and it was not qualified by any higher level interactions with *item-novelty* ( $\beta = -0.14$ ,  $SE = 0.22$ ,  $z = -0.64$ ,  $p$

= 0.52). This suggested no evidence that the extent to which participants improved on this task between pre and post-test differed according to *variability-conditions*, or that this differed for trained versus untrained items.

Although not part of the key predictions, the analysis also explored 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 ( $\beta = -0.31$ ,  $SE = 0.14$ ,  $z = -2.28$ ,  $p = .02$ ) while the contrast between “neutral” and “hard” was not ( $\beta = 0.07$ ,  $SE = 0.13$ ,  $z = 0.55$ ,  $p = .58$ ). This was due to the fact that there was improvement for “neutral” ( $M_{pre} = 0.56$ ,  $SD_{pre} = 0.14$ ,  $M_{post} = 0.64$ ,  $SD_{post} = 0.15$ ) and “hard” trials ( $M_{pre} = 0.53$ ,  $SD_{pre} = 0.16$ ,  $M_{post} = 0.63$ ,  $SD_{post} = 0.15$ ) but not for “easy” trials ( $M_{pre} = 0.66$ ,  $SD_{pre} = 0.16$ ,  $M_{post} = 0.67$ ,  $SD_{post} = 0.16$ ).

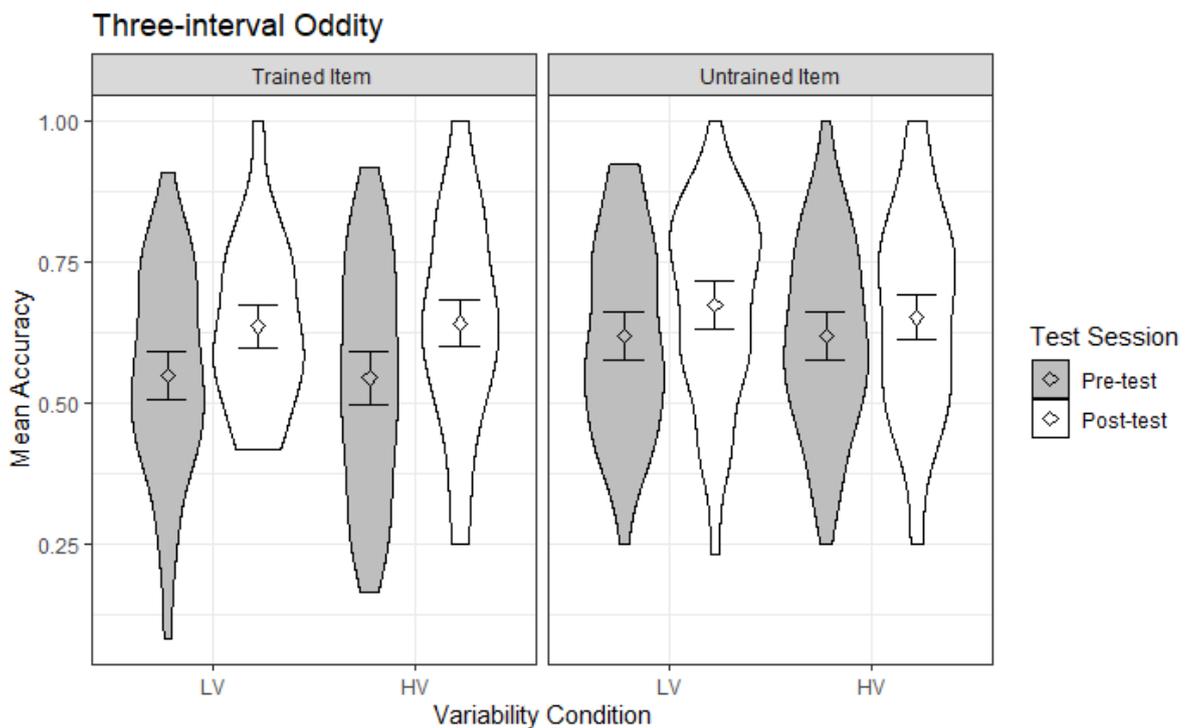


Figure 8 Mean proportion of correct in Three Interval Oddity task for LV (Low Variability) and HV (High Variability) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals.

#### 2.3.4.2 Picture Identification

The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were the factor *voice-novelty* (trained voice, untrained voice) and the factor *variability-condition* which had two levels (LV, HV). The mean accuracy is displayed in Figure 9.

There was a main effect of *voice-novelty* ( $\beta = 1.28$ ,  $SE = 0.21$ ,  $z = 6.21$ ,  $p < .001$ ) reflecting higher performance in trials with trained voices. Participants in the LV group performed better than those in the HV group ( $\beta = -0.71$ ,  $SE = 0.31$ ,  $z = -2.32$ ,  $p = 0.02$ ) and there was a significant interaction between *voice-novelty* and *variability-condition* ( $\beta = -1.20$ ,  $SE = 0.36$ ,  $z = -3.37$ ,  $p < .01$ ). Breaking this down by *variability-condition*: for each condition there was significantly better performance with trained than untrained voices (LV:  $\beta = 1.88$ ,  $SE = 0.30$ ,  $z = 6.18$ ,  $p < 0.001$ ; HV:  $\beta = 0.68$ ,  $SE = 0.24$ ,  $z = 2.86$ ,  $p < 0.01$ ), indicating greater ease with the familiar voice. Breaking down by *voice-novelty*: For the trained voice, performance was higher in the LV condition than in the HV condition ( $\beta = -1.32$ ,  $SE = 0.44$ ,  $z = -2.97$ ,  $p < 0.01$ ). Importantly, for untrained voices, the difference between conditions was not significant ( $\beta = -0.11$ ,  $SE = 0.24$ ,  $z = -0.48$ ,  $p = 0.64$ ), indicating no evidence for greater generalisation following high variability training.

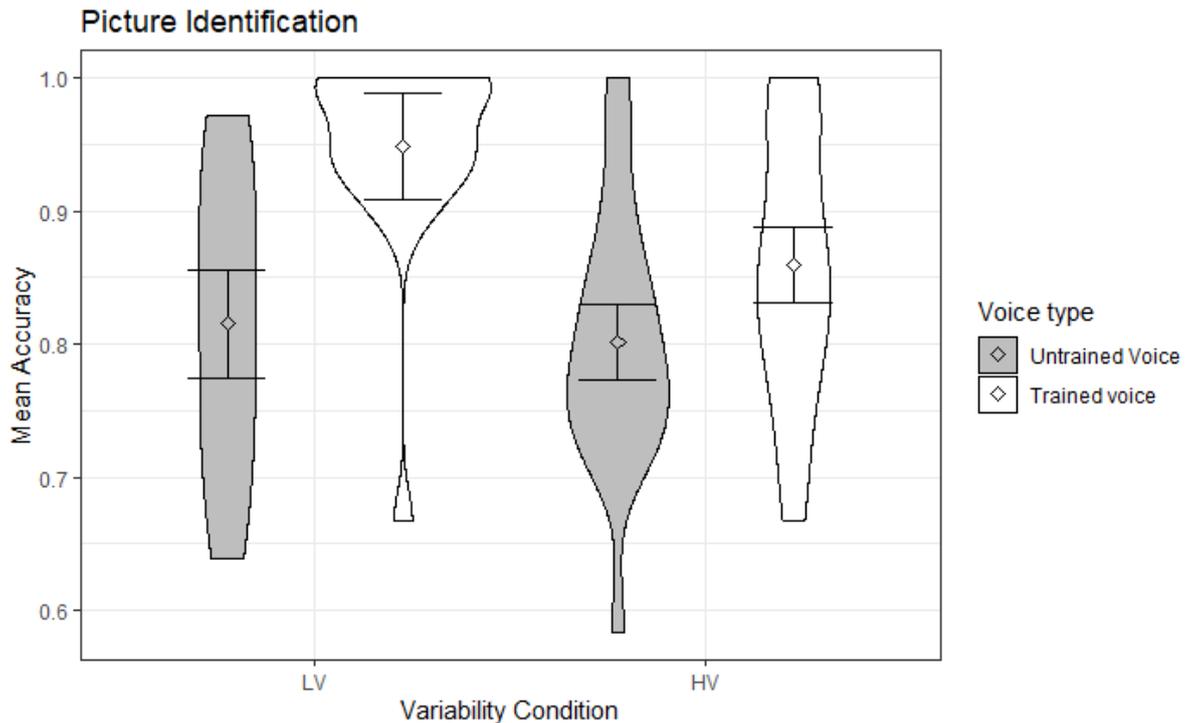


Figure 9 Mean proportion of correct of Picture Identification for LV (Low Variability) and HV (High Variability) training groups for untrained voices and trained voices. Error bars show 95% confidence intervals.

### 2.3.5 Production tests

#### 2.3.5.1 Coding and inter-rater reliability analyses

The same methods were used for both production tests. 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 myself and Rater 2 was a female native Mandarin speaker 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.

If there was no sound or the tone was unrecognizable, the rater coded 0 when identifying the tone. Data from these trials were removed from the dataset before analyses were conducted. In addition, all of the data from one participant was removed from the analyses due to bad recording quality resulting from a technical error. In total, this resulted in 3.54% (251/7080) of production trials being removed from analysis (*Word Repetition*: Pre-test 1.06% (30/2832); Post-test 4.34% (123/2832); *Picture Naming* 6.92% (98/1416)). 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 in 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 all three 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  $kappa = 0.41$  (“moderate agreement”), and for Pinyin accuracy  $kappa = 0.35$  (“fair agreement”; Landis & Koch, 1977). For the Picture Naming test, for Tone accuracy  $kappa = 0.67$  (“substantial agreement”) and for Pinyin accuracy  $kappa = 0.56$  (“moderate agreement”); For the Tone rating, the package “irr” in R was used to assess the intra-class correlation (McGraw & Wong, 1996) based on an average-measures, two-way mixed-effects model. For Word Repetition,  $ICC = 0.22$  and for Picture Identification  $ICC = 0.34$ ; according to Cicchetti (1994), values less than

.40 are regarded as “poor”. Given this, I do not include analyses with Tone Rating as the dependent variable (though these data are included in the data set [https://osf.io/j6s7w/?view\\_only=497e0e8ee7ff4e7387984690eafd4b5a](https://osf.io/j6s7w/?view_only=497e0e8ee7ff4e7387984690eafd4b5a)). All of the analyses presented in Sections 2.3.5.2 and 2.3.5.3 were based on Rater 2 (the naive rater).

### 2.3.5.2 Word Repetition

#### 2.3.5.2.1 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, HV) and *item-novelty* (trained, untrained). The mean accuracy, split by test-session and training condition, is shown in Figure 10.

At pre-test, there was no significant difference between the LV and the HV group ( $\beta = 0.01$ ,  $SE = 0.20$ ,  $z = 0.06$ ,  $p = 0.95$ ) suggesting the groups started at a similar level. There was also no difference between trained and untrained words at pre-test ( $\beta = -0.06$ ,  $SE = 0.09$ ,  $z = -0.73$ ,  $p = 0.47$ ).

For both groups, participants’ performance increased significantly after training ( $M_{pre} = 0.71$ ,  $SD_{pre} = 0.09$ ,  $M_{post} = 0.78$ ,  $SD_{post} = 0.09$ ,  $\beta = 0.41$ ,  $SE = 0.10$ ,  $z = 4.17$ ,  $p < .001$ ). There was also an *item-novelty* by *test-session* interaction ( $\beta = 0.26$ ,  $SE = 0.13$ ,  $z = 2.02$ ,  $p = 0.04$ ). Breaking down by *item-novelty*, the analysis suggested that the improvement was bigger for untrained items ( $M_{pre} = 0.70$ ,  $SD_{pre} = 0.08$ ,  $M_{post} = 0.80$ ,  $SD_{post} = 0.09$ ,  $\beta = 0.54$ ,  $SE = 0.12$ ,  $z = 4.53$ ,  $p < .001$ ) than trained items ( $M_{pre} = 0.71$ ,  $SD_{pre} = 0.09$ ,  $M_{post} = 0.77$ ,  $SD_{post} = 0.08$ ,  $\beta = 0.28$ ,  $SE = 0.12$ ,  $z = 2.41$ ,  $p = 0.02$ ). Critically, the interaction between *variability-condition* and *test-session* was not significant ( $\beta = 0.06$ ,  $SE = 0.11$ ,  $z = 0.47$ ,  $p = 0.64$ ), and they were not qualified by a higher level interactions with *item-novelty* ( $\beta = 0.11$ ,  $SE = 0.22$ ,  $z = 0.50$ ,  $p = .62$ ). This suggests there is no evidence that participants’ improvement in their

production of tones was affected by their *variability-condition*, or that this differed for *trained* versus *untrained* items.

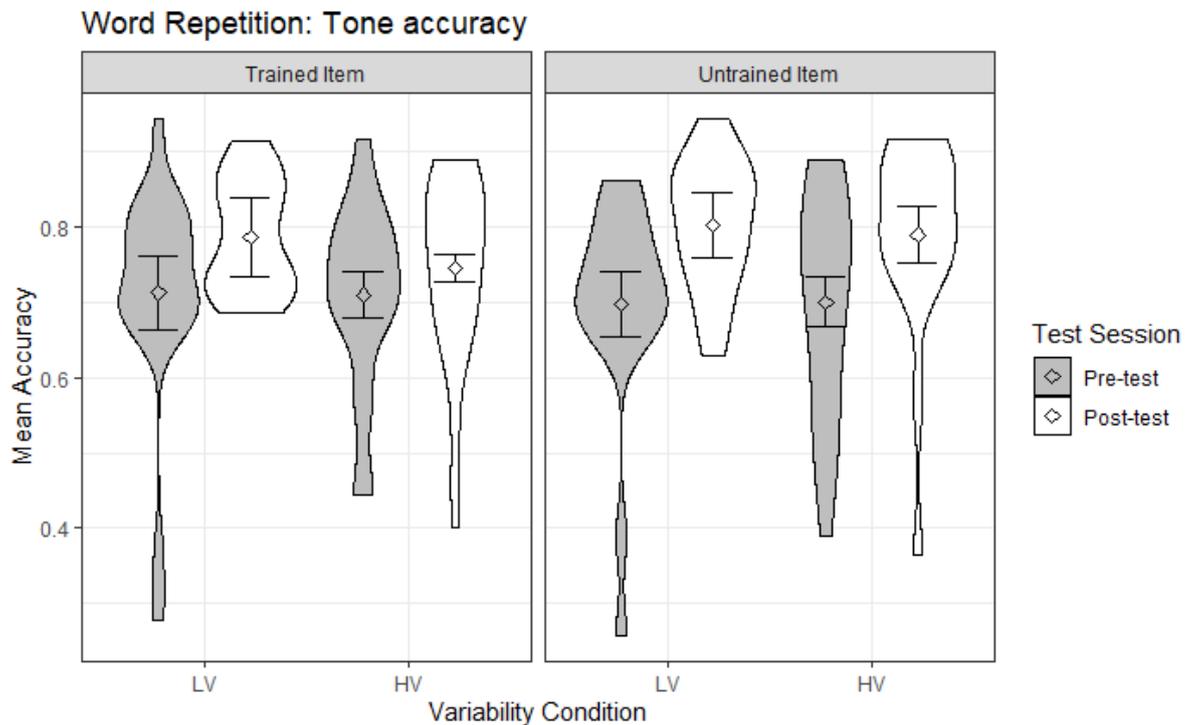


Figure 10 Accuracy of Word Repetition for LV (Low Variability) and High Variability (HV) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals.

### 2.3.5.2.2 Pinyin accuracy

The predicted variable was whether the participants produced the correct string of phonemes (1/0) in each trial (as determined by Rater 2). The predictors were *test-session* (pre-test, post-test), *variability-condition* (LV, HV) and *item-novelty* (trained, untrained). Mean pinyin accuracy is displayed in Figure 11.

At pre-test, there was no significant difference between the LV and the HV group ( $\beta = -0.01, SE = 0.12, z = -0.10, p = 0.92$ ) suggesting that the groups started at a similar level. There was also no difference between *trained* and *untrained* words at pre-test ( $\beta = 0.15, SE = 0.09, z = 1.75, p = 0.08$ ).

Participants showed significant improvement after training ( $M_{pre} = 0.55$ ,  $SD_{pre} = 0.10$ ,  $M_{post} = 0.59$ ,  $SD_{post} = 0.09$ ,  $\beta = 0.20$ ,  $SE = 0.06$ ,  $z = 3.65$ ,  $p < .001$ ). However, there was no *test-session* by *item-novelty* interaction ( $\beta = 0.12$ ,  $SE = 0.11$ ,  $z = 1.07$ ,  $p = 0.29$ ). There was also no evidence that different variability conditions resulted in different amounts of improvement (*test-session* by *variability-condition*:  $\beta = 0.05$ ,  $SE = 0.11$ ,  $z = 0.46$ ,  $p = .65$ ) or any interaction between *variability condition*, *test-session* and *item-novelty* ( $\beta = 0.11$ ,  $SE = 0.22$ ,  $z = 0.50$ ,  $p = .62$ ). This suggests there is no evidence that participants' improvement in pinyin accuracy was affected by their *variability-condition*, or that this differed for trained versus untrained items.

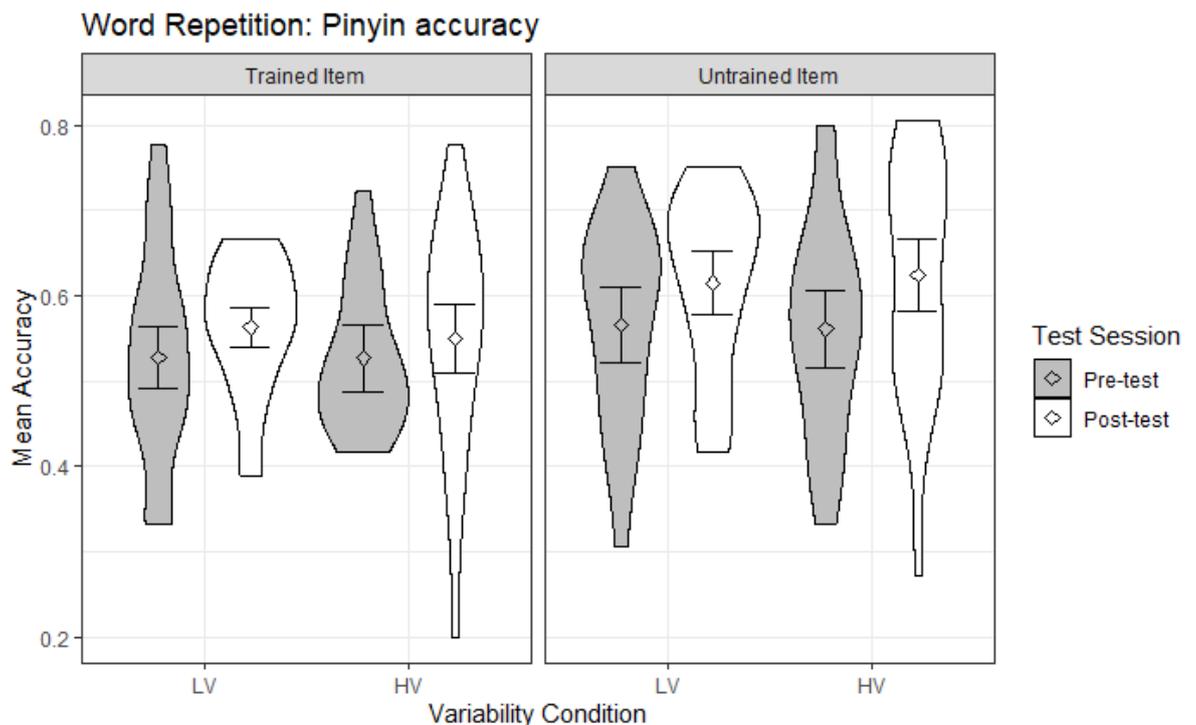


Figure 11 Mean pinyin accuracy of Word Repetition for LV (Low Variability) and HV (High Variability) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals.

### 2.3.5.3 Picture Naming

#### 2.3.5.3.1 Tone accuracy

The predicted variable was whether a correct response was given (1/0) on each trial (as identified by the coder). There was only one predictor, *variability-condition* (LV, HV) for both models. The descriptive statistics are displayed in Figure 12.

Participants in the HV group did not outperform those in the LV group (with the means in the reverse direction;  $\beta = -0.34$   $SE = 0.19$ ,  $z = -1.83$ ,  $p = 0.07$ ). This suggests there is no evidence that participants' ability to produce the tones accurately differed according to their *variability-condition*.

#### 2.3.5.3.2 Pinyin Accuracy

The predicted variable was whether the participants produced the correct string of phonemes (1/0) in each trial and there was a single predictor *variability-condition* (LV, HV). For both models there was no significant difference between variability conditions ( $\beta = 0.09$ ,  $SE = 0.22$ ,  $z = 0.42$ ,  $p = 0.68$ ).

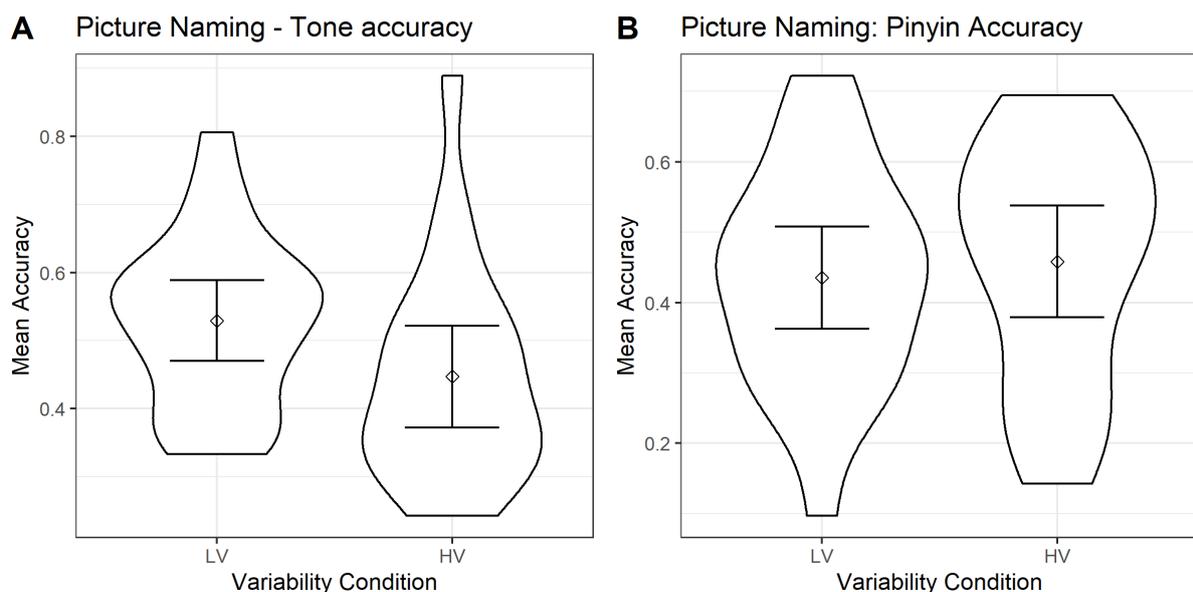


Figure 12 Tone accuracy and Pinyin accuracy of Picture Naming for LV (Low Variability) and HV (High Variability) training groups. Error bars show 95% confidence intervals.

### 2.3.6 Analyses with Individual Aptitude

In order to look at the effect of learner aptitude and the interaction between this factor and variability condition, the mean accuracy at pre-test on the Pitch Contour Perception Test for each participant and the slope at pre-test on Categorization of Synthesized Tonal Continua were calculated. These scores were centered and each was used as a continuous predictor (*aptitude*, *the Pitch Contour Perception Test* and *Categorisation of Synthesized Tonal Continua*) and added to each of the models reported above. In addition, the interactions between these factor and key experimental factors were also added (see Table 6: *Pitch Contour Perception Test*; Table 7: *Categorisation of Synthesized Tonal Continua*). Based on Perrachione et al. (2011) and Sadakata and McQueen (2014), for the measures of tone-learning, high variability should benefit high aptitude participants more, while low variability would benefit low aptitude participants more. In the current design, a continuous measure of individual ability, rather than a binary division of high and low aptitude, was used. Therefore, it was predicted that there would be a stronger positive correlation between *aptitude* and amount of learning in the high variability condition compared with the low variability condition. In the tests administered only post training (i.e. Picture Identification and Picture Naming) this would show up as an interaction between *aptitude* and *variability-condition*. In the models for the pre- and post-test data (i.e. Three Interval Oddity and Word Repetition) this would show up as a three-way interaction between *variability-condition*, *test-session* and *aptitude*. Where relevant, the models also included the interactions between these factors and *voice-novelty* (Picture Identification) and *item-novelty* (Three Interval Oddity and Word Repetition). Note that there are no clear directional hypotheses here: Perrachione et al. (2011) found the interaction in a test with untrained voices and trained items, and Sadakata and McQueen (2014) found the interaction in a test with trained voices and trained items. For

training, in principal both the two-way interaction of *aptitude* by *variability-condition* and the three-way interaction of *aptitude* by *variability-condition* by *training-session* are of interest.

Each model reported in Table 6 and Table 7 contained all the fixed effects included in the original models in addition to the fixed effects listed in the table (note that to avoid convergence issues due to over complex models, the analyses did *not* attempt to include the complete set of interactions for every combination of experimental variables with aptitude – only those relevant to the hypotheses). The original analyses attempted to have a full random effects structure for these fixed effects, however in some cases it was necessary to remove correlations between slopes due to problems with convergence (for details, see [https://osf.io/j6s7w/?view\\_only=497e0e8ee7ff4e7387984690eafd4b5a](https://osf.io/j6s7w/?view_only=497e0e8ee7ff4e7387984690eafd4b5a)). Note that the analyses did not include models for the pinyin measures, since the measures of aptitude were relevant to tone learning only.

The results for the *Pitch Contour Perception Test* as the aptitude measure are shown in Table 6 and the results for the *Categorisation of Synthesized Tonal Continua* as the aptitude measure are shown in Table 7. Table 6 suggested the *Pitch Contour Perception Test* is a positive predictor of performance in each of the tests and in training, however, there was no interaction with 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 the hypothesis - did this differ by training condition (no interaction with *variability-condition* or with *test-session* by *condition*). As can be seen from Table 7, *Categorisation of Synthesized Tonal Continua* was not predictive in any instance, suggesting this might not be a good predictor of ability with lexical tones. In sum, participants with higher ability on the *Pitch Contour Perception Test* were better at the experimental 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.

Table 6 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.

Data Set	Coefficient Name	Statistics
<i>Word Repetition: Tone Accuracy (Pre/Post)</i>	<b>Aptitude</b>	<b><math>\beta = 0.10</math>, SE = 0.04, z = 2.44, p = .015</b>
	Aptitude by <i>Test-Session</i>	$\beta = -0.001$ , SE = 0.06, z = -0.02, p = .986
	Aptitude by <i>variability-condition</i>	$\beta = 0.01$ , SE = 0.06, z = 0.21, p = .833
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i>	$\beta = 0.03$ , SE = 0.07, z = .52, p = .603
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i> by <i>Item-Novelty</i>	$\beta = -0.21$ , SE = 0.12, z = -1.70, p = .089
<i>Three Interval Oddity (Pre/Post)</i>	Aptitude	$\beta = 0.05$ , SE = 0.05, z = 1.13, p = .257
	Aptitude by <i>Test-Session</i>	$\beta = 0.0001$ , SE = 0.03, z = 0.03, p = .978
	Aptitude by <i>variability-condition</i>	$\beta = -0.09$ , SE = 0.09, z = -1.00, p = .318
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i>	$\beta = 0.05$ , SE = 0.07, z = 0.75, p = .452
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i> by <i>Item-Novelty</i>	$\beta = -0.09$ , SE = 0.13, z = -0.74, p = .459
<i>Training</i>	<b>Aptitude</b>	<b><math>\beta = 0.12</math>, SE = 0.06, z = 1.99, p = .047</b>
	Aptitude by <i>variability-condition</i>	$\beta = -0.02$ , SE = 0.06, z = -0.312, p = .755
<i>Picture Identification (Post Only)</i>	Aptitude	$\beta = 0.142$ , SE = 0.09, z = 1.53, p = .127
	Aptitude by Voice Novelty	$\beta = 0.08$ , SE = 0.12, z = 0.67, p = .504
	Aptitude by <i>variability-condition</i>	$\beta = -0.02$ , SE = 0.18, z = -0.10, p = .921
	Aptitude by <i>variability-condition</i> by <i>Voice-Novelty</i>	$\beta = 0.35$ , SE = 0.22, z = 1.55, p = .122
<i>Picture Naming: Tone Accuracy</i>	Aptitude	$\beta = 0.03$ , SE = 0.05, z = 0.51, p = 0.614
	Aptitude by <i>variability-condition</i>	$\beta = -0.09$ , SE = 0.11, z = -0.81, p = .416

Table 7 Statistics obtained when adding in participant aptitude (as measured by performance on the Categorisation of Synthesized Tonal Continua task at pre-test) into the models predicting performance on the test and training tasks.

Data Set	Coefficient Name	Statistics
<i>Word Repetition: Tone Accuracy (Pre/Post)</i>	Aptitude	$\beta = 0.04$ , SE = 0.07, z = 0.51, p = .609
	Aptitude by <i>Test-Session</i>	$\beta = -0.03$ , SE = 0.10, z = -0.32, p = .751
	Aptitude by <i>variability-condition</i>	$\beta = -0.17$ , SE = 0.14, z = -1.24, p = .216
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i>	$\beta = -0.17$ , SE = 0.18, z = -1.33, p = .184
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i> by <i>Item-Novelty</i>	$\beta = 0.01$ , SE = 0.23, z = 0.05, p = .960
<i>Three Interval Oddity (Pre/Post)</i>	Aptitude	$\beta = 0.08$ , SE = 0.07, z = 1.27, p = .203
	Aptitude by <i>Test-Session</i>	$\beta = -0.02$ , SE = 0.05, z = -0.47, p = .639
	Aptitude by <i>variability-condition</i>	$\beta = -0.03$ , SE = 0.14, z = -0.20, p = .844
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i>	$\beta = -0.003$ , SE = 0.10, z = -0.03, p = .978
	Aptitude by <i>variability-condition</i> by <i>Test-Session</i> by <i>Item-Novelty</i>	$\beta = -0.22$ , SE = 0.20, z = -1.10, p = .269
<i>Training</i>	Aptitude	$\beta = 0.16$ , SE = 0.10, z = 1.58, p = .115
	Aptitude by <i>variability-condition</i>	$\beta = -0.11$ , SE = 0.10, z = -1.07, p = .286
<i>Picture Identification (Post Only)</i>	Aptitude	$\beta = 0.17$ , SE = 0.14, z = 1.26, p = .209
	Aptitude by Voice Novelty	$\beta = -0.27$ , SE = 0.27, z = -0.99, p = .322
	Aptitude by <i>variability-condition</i>	$\beta = -0.02$ , SE = 0.17, z = -0.14, p = .892
	Aptitude by <i>variability-condition</i> by <i>Voice-Novelty</i>	$\beta = 0.01$ , SE = 0.33, z = 0.02, p = .982
<i>Picture Naming: Tone Accuracy</i>	Aptitude	$\beta = 0.06$ , SE = 0.10, z = 0.59, p = 0.553
	Aptitude by <i>variability-condition</i>	$\beta = -0.12$ , SE = 0.20, z = -0.58, p = .564

## 2.4 Discussion

This chapter reported an experiment directly contrasting multi-speaker (HV) and single speaker (LV) phonetic training for Mandarin lexical tone contrasts for English-speaking individuals. In contrast to previous work, the full set of four Mandarin tones was included, imbedded in real Mandarin words. The key aim was to explore interactions between *variability condition* and individual aptitude. Two specific tests of individual aptitude were conducted:

The Pitch Contrast Perception Test (following Perrachione et al. 2011) and the Categorisation of Synthesised Tonal Continua, following Sadakata and McQueen (2014). Benefits of training were assessed by a battery of perception and production measures, some of which were conducted both pre- and post- training. Generalisation was measured with the use of novel voices (included in all tests) and with novel items (included in in the pre- to post- tests). This discussion will critically evaluate the findings in each task in turn.

#### 2.4.1 *Pitch Contour Perception Test & Categorisation of Synthesised Tonal Continua*

Although the primary goal of these tests was to provide a measure of participants' baseline aptitude, the tests were conducted both at pre- and post- test. Starting with PCPT, it should be noted that Perrachione et al. (2011) did not actually conduct this test at post-test. However, interestingly, analyses here (Section 2.3.2.1) demonstrated that performance in this test improved from pre- to post- training. This suggests that this measure is not a "pure" measure of individual differences since it also appears to be affected by learning experience. This is not too surprising as this task involves measuring participants' ability to identify Mandarin tones, and this should be benefited by training to use the tones in word identification. Given this finding only participants' *pre-test* scores were used as the measure of aptitude in subsequent analyses. As for Sadakata and McQueen's (2014) measure of categorisation ability of synthesised Mandarin tones, they did not find any difference from pre- to post- tests thus used combined data from pre- and post- test to compute participants slopes. Although I also did not find any improvement from pre- to post-test, it appears that this test may not have been a good measure of aptitude for our participants: The majority of participants failed to meet the slope threshold used in Sadakata and McQueen. Moreover, it was found that most of the participants were unable to consistently categorise the end points of the continua. Looking at the mean accuracy, it can be see that their pre-test accuracy in identifying the end points of the continuum is around 50% chance level ( $M_{HV} = 59\%$ ,  $M_{LV} = 64\%$ ). Furthermore, although a

significant correlation between the *Pitch Contour Perception Test* and the *Categorisation of Synthesized Tonal Continua* measure was found, *Categorisation of Synthesized Tonal Continua* was not predictive of any of the performance measures even at pre-test, or of training, whereas the *Pitch Contour Perception Test* was (as further discussed below in each section). It is unclear why the current results with the *Categorisation of Synthesized Tonal Continua* differ from those of Sadakata & McQueen (2014), since the test was constructed to replicate their test.

#### 2.4.2 Performance in Training and Picture Identification

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 multiple speakers HV training. This difference was present from the first training session and increased over time. Greater difficulty with multiple speaker input is in line with the findings of Perrachione 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 adaptation to speaker specific cues, whereas in the HV conditions participants have to adapt to multiple speakers. An additional difficulty in the current HV condition is that speakers change randomly trial-by-trial, requiring constant adaptation, which may be effortful for participants (Magnuson & Nusbaum, 2007). I return to the point below. Importantly, however, for both 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. Turning to the role of learner aptitude in this task (as measured by performance on the *Pitch Contour Perception Test* and *Categorisation of Synthesized Tonal Continua* at pre-test). Overall, *Pitch Contour Perception Test* was found to be a significant predictor ( $p = 0.047$ ) of

performance during training, while the results for *Categorisation of Synthesized Tonal Continua* was not predictive on any occasion. As discussed above, this suggested that *Pitch Contour Perception Test* was sufficiently sensitive to capture the ability to discriminate Mandarin tones, as required in the current training task. However, critically there was no evidence for an interaction with training condition for either *aptitude* measure. The *Picture Identification* test – a version of the training task without feedback which was administered post-training – replicated the 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 worse with *untrained speakers* than with *trained speakers*, but were above chance. This indicates across speaker generalisation which did *not* depend on speaker variability in training.

#### 2.4.3 *Three Interval Oddity Task*

The Three Interval Oddity task was administered at both pre- and post-test. It had not been used in the previous studies, but allowed us to use a pre- to post - test design, and also to look at participants' performance with both *untrained items* and *untrained speakers*. Again, the results indicates that participants improved in their discrimination of tones following training: (58% performance prior to training, 65% post training). There was also evidence of generalisation across both voices (since novel speakers were used for all of the test items) and items, as they improved in their ability to discriminate between minimal trio items even for *untrained items*. There were no differences in the extent of improvement for *trained* compared with *untrained* items, indicating that their improved tone discrimination ability was not item specific. Critically, this improvement following training occurred equally across the both variability conditions, indicating that input variability was not necessary for generalisation. Returning the key prediction, that high aptitude participants would benefit more from high variability training, there was no evidence of the relevant interaction for either of the aptitude

measures. This was despite the fact that the *Pitch Contour Perception Test* measure was predictive of performance in this task at pre-test.

Another finding from the Three Interval oddity test that is worth noting, although it does not concern the hypotheses directly, is that some trial types were harder than others. Recall that this test involved participants hearing three different stimuli each produced by a different speaker, which makes noting the similarity across two of the stimuli much harder – this was also discovered in pilot work, where even before training participants were near ceiling with an equivalent task in which the same speaker produced all three stimuli within a single trial. However, analyses of *trial-type* demonstrated that participants were additionally affected by the gender of the three speakers producing each of the stimuli. Specifically, at pre-test, participants showed best performance for trials where one of the speakers was male and the other two were female, and the target “odd one” was the male speaker (“*easy*” trials). In contrast, they showed worst performance if there was one male and two female speakers, but the “odd one” 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, the analyses showed that performance only increased for the trials where the odd one was not the lone male (the “*neutral*” and “*hard*” ones), but not for those where the male was the odd man. Given that participants are not near ceiling at pre-test (58%), 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 realise are less reliable, masking improvement based on tone for these particular test items.

#### 2.4.4 *Word Repetition & Picture Naming*

In this study, two production tasks were used: A Word Repetition task administered pre- and post- training, in which participants repeated back Mandarin words, and a Picture Naming task testing vocabulary recall, which was administered at post-test only. High variability perceptual training for tones has been previously found to transfer to production (Bradlow and Pisoni, 1999; Zeromskaite, 2014), however the benefits of high variability and low variability training have not been contrasted.

In the Word Repetition task, there was 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: 71% to 78%) and in the Picture Naming task, participants showed an ability to recall and reproduce the correct tone, although unsurprisingly with less accuracy than in the repetition task (53%). For Word Repetition, the analyses also looked at transfer to untrained words: As in the perception tasks, there was once again equivalent improvement for both trained and untrained items. Together, these results provide evidence that purely perceptual training on tone contrast can transfer to production, as well as to novel items.

In addition to looking at the production of *tones*, the current study also looked at participants' ability to produce the correct segmental phonology (Pinyin-score). Participants showed a small but significant improvement on this measure in Word Repetition (55% correct at pre-test, 59% at post-test), and some ability to recall the segments in the Picture Naming test (44% correct). This indicates some learning of segmental phonology due to training, despite the fact that the focus of the training task was on training tonal information through the presentation of tonal minimal-pairs.

Turning to the role of variability, the predicted benefit of high variability training was *not* evident in any of the measures in either of the production tasks. With regard to aptitude, although performance on the *Pitch Contour Perception Test* at pre-test was predictive of participants' ability to produce tones in both tasks (indicating a relationship between participants' perceptual and production ability), there was no predicted interaction between aptitude and *variability-condition* in either task. Again, *Categorisation of Synthesized Tonal Continua* was not predictive even of performance at pre-test, suggesting that this measure is not a good representation of aptitude for tone discrimination.

#### 2.4.5 Evaluation

The current study did not find support either for a general advantage of HV for generalisation, as reported in the phonetic training literature (e.g. Logan et al., 1991; Lively, et al., 1993), nor for an interaction between input variability and individual aptitude as shown in the recent work of Perrachione et al. (2011) and Sadakata and McQueen (2014).

This is discussed in more depth in the next chapter, however here it should be noted that there are a variety of differences across the studies which in general tend to *increase* the difficulty of the tasks in the current study compared with previous research. The *Pitch Contour Perception Test* adapted from Perrachione et al. (2011) used all six Mandarin main vowels (where they used five, without /u/) and all four Mandarin tones (where they did not use Tone 3). In addition, Three Interval Oddity task also included all four tone contrasts, including those involving T3 (which again Perrachione et al. (2011) did not use). Most importantly, in Training, real Mandarin words were used and I also trained participants with all possible Mandarin tonal contrasts including Tone 3. Tone 3 was considered perceptually the most confusable tone for L2 learners (Dong, Tsubota & Dantsuji, 2013; Hao, 2014). Whilst it is not clear why these increases in complexity should remove the interactions with aptitude (if anything it should be

expected that increased task difficulty increases the range of scores and make it easier to find effects of individual differences) it does make it harder to contrast the current results with those of the previous studies.

Another major difference between the current study and many previous HVPT studies is that the HV condition used a design in which the speakers in training were randomly intermixed. As pointed out in section 2.4.2, this requires trial-by-trial adaptation to each speaker, which was not needed in the corresponding single speaker LV conditions. This may place a burden on learners and may at least partially account for why there was such reduced performance in training in the HV condition compared with the LV condition. Early work (Nusbaum & Morin, 1992) has suggested that attending to different speakers is cognitive-demanding even for L1 processing, which is reflected in increased RT. Thus, using intermixed materials in a HV design may increase the overall cognitive load thus impairing the training results (Mattys & Wiget, 2011). In contrast, most HVPT studies in the literature present stimuli from the same speaker within a single block (e.g. Iverson, Hazan & Bannister, 2005; Logan et al. 1991). Sadakata & McQueen (2014) used this type of blocked design. The original study by Perrachione et al. (2011) didn't intermix speakers. However, they conducted a second blocked version of the HV condition and found that this improved performance during the training task compared with HV unblocked training. It may be that there is an interaction between the potential benefits of exposure to cues from multiple speakers, and the added complexity of dealing with constantly varying speakers during training. If this is the case, in the current more complex training paradigm, in order to see benefits of multi-speaker input, even if only occurring for relatively high aptitude participants, it might be necessary to remove the increased complexity caused by trial-by-trial inconsistency in speakers. I tested this by adding a third high variability blocked (HVB) condition to the experiment, using identical

stimuli to the current HV condition but with speakers presented in separate blocks. The results from this condition are compared with those of the previous two conditions and these are presented as Study 2 in the next chapter.

### 3. Study 2

#### 3.1 Introduction

The previous chapter reported a study comparing two types of computerised phonetic training on Mandarin lexical tones: multi-speaker (HV) and single-speaker (LV). In contrast to previous literature, there was no advantage of HV training on generalisation and there was also no interaction between variability condition and participants' aptitude, as measured with the *Pitch Contour Perception Test & Categorisation of Synthesized Tonal Continua*. However a potential factor that might reduce performance in the HV condition in previous experiment is the fact that the four speakers used training varied trial by trial. The current experiment explored whether removing this trial by trial inconsistency by introducing blocked training, might either allow an HV benefit in generalisation to emerge, or might reveal a relationship between variability and individual difference in training.

As discussed in the last chapter, Perrachione et al. (2011) compared high variability conditions in which speakers were either intermixed or blocked, finding a benefit of blocking particularly for low aptitude learners. The current study, following Perrachione et al. (2011), introduces a third condition: high variability stimuli blocked by speaker (HVB). Procedure and stimuli were identical to the previous HV training except that the speakers were presented in blocks. Note that the choice to manipulate only speaker-variability means that the HVB condition is matched to the LV condition in terms of trial-by-trial inconsistency (i.e. the amount that the Mandarin words used in training change trial by trial), unlike in Sadakata and McQueen (2014) where, even though they blocked by speaker, the high variability condition contained more trial-by-trial variability in terms of items. This high variability blocked condition (HVB) is compared to the previous two conditions- i.e. high variability not blocked (HV) and low variability (LV). It is predicted that this condition will be easier for participants than the

previous HV condition which might allow a benefit of speaker variability on generalisation to emerge.

The current chapter presents data from this new condition, repeating the analyses conducted in the previous chapter but replacing the contrast between the LV and HV conditions with the contrast between this new HVB condition and each of the two previous conditions. In addition to these analyses, which use frequentist methods ( $p$ -values), this chapter also presents a second set of analyses using *Bayes Factors*. The Bayes factor statistics can be used to assess the strength of evidence for one hypothesis (H1) over another (the null hypothesis). It has the key advantage over  $p$ -values for interpreting null results: Non-significant result using frequentist statistics (e.g. where  $p > .05$ ) does *not* tell whether there is evidence for the null, as opposed to no evidence for any conclusion at all, or even evidence against the null. This means that where there is no evidence for the hypotheses about variability, it is inappropriate to reduce the confidence in these hypotheses on the basis of the fact that  $p > .05$  (despite the fact that reducing confidence in a theory following non-significant results is common practice, see Dienes (2014) for a discussion). In contrast, Bayes Factors allow us to differentiate three situations with regard to the evidence: the situation where there is substantial evidence for the hypothesis (compared with the null), the situation where there is substantial evidence for the null (compared with the hypothesis) and the situation where the evidence is ambiguous.

## 3.2 *Method*

### 3.2.1 *Participants*

Twenty extra adults recruited from UCL Psychology Subject Pool participated in the experiment, they were assigned to the HVB condition. Updated participant information can be seen in Table 8. There was no difference between these groups in age,  $F(2, 57) = 1.95$ ,  $p =$

0.15. Again, all Participants were native English speakers and they had no known hearing, speech, or language impairments.

*Table 8 Mean age, age range, average number of languages learned and mean starting age of learning the first L2 for participants in the high variability blocked condition.*

Condition	Mean Age	Age Range	Languages	Average
			Learned	Starting Age
High Variability blocked	22.05 (1.4)	19-30	2.0 (1.3)	11.8 (0.4)

### 3.2.2 Stimuli & Procedure

In the new HVB training design, the stimuli in training were identical to those in the previous HV condition (i.e. 18 picture/word pairs, each word used as the target word four times, resulting in 288 trials with 4 speakers). However, from Day 1 to Day 4 of training (i.e., Session 2-5), only one speaker was used on each training session. On Days 5 and 6 of training (i.e., Sessions 6 and 7), participants heard all four speakers, each in a separate block, with each word was repeated twice in each voice on these days (day 5 and 6 were identical for each participant). The order in which speakers were used was rotated across participants, although the “trained” speaker that was used in the test tasks always occurred on Day 3 (i.e., Session 4) and was used in the third block on days 5 and 6 (Table 9). After each block, the number of coins they had earned so far was displayed on the screen.

Other tests and procedures were the same as in the previous study (Section 2.2).

*Table 9 Counterbalancing of voices for High variability blocked design and Picture Identification.*

Task	Training Day	Voice				
		Version 1	Version 2	Version 3	Version 4	Version 5
Training HVB	Day1	M1	M1	F1	F2	F3
	Day2	F3	F1	M2	F1	F2
	Day3	F1	2	F3	M1	M2

	Day4	M2	M2	F2	F3	M1
	Day5	All	All	All	All	All
	Day6	All	All	All	All	All
Picture Identification						
<i>Trained voice</i>		F1	F2	F3	M1	M2
<i>Untrained voice</i>		F2	F3	M1	M2	F1

### 3.3 Results

#### 3.3.1 Statistical Approach

The analytic procedure remained the same as for the previous study (Section 2.3.1). The only change is for the coding of the factor *variability-condition*. In each of the analyses, the factor *variability-condition* now has three levels (low variability [LV], high variability [HV], and high variability blocked [HVB]). These were coded into two contrasts with HVB as the baseline (HVB versus LV, HVB versus HV). An exception to this is the training data, where a model containing all three conditions would not converge and a different approach was taken, as described in Section 3.3.3. The models also included the interactions between these contrasts and the other factors. Centered coding was used which ensured that other effects were evaluated as averaged over all three levels of *variability-condition* (rather than the reference level of LV<sup>4</sup>). 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/j6s7w/?view\\_only=497e0e8ee7ff4e7387984690eafd4b5a](https://osf.io/j6s7w/?view_only=497e0e8ee7ff4e7387984690eafd4b5a)

In addition to the frequentist analyses, in order to aid interpretation of key null results Bayes factor analyses were also performed. The approach for these is described within the relevant section (Section 3.3.7).

---

<sup>4</sup> 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.

### 3.3.2 Individual Aptitude Tasks

#### 3.3.2.1 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 *variability-conditions* (LV versus HVB; HV versus HVB) and *test-session* (pre-test, post-test). There was no significant difference between the LV and HVB groups ( $\beta = -0.17$ ,  $SE = 0.26$ ,  $z = -0.66$ ,  $p = 0.51$ ). However, there was a significant difference between the HV and HVB groups ( $\beta = -0.52$ ,  $SE = 0.26$ ,  $z = -2.05$ ,  $p = 0.04$ ) at pre-test on this measure, suggesting the newly recruited HVB group was naturally better than the HV group. Participants showed significant improvement after training ( $\beta = 0.21$ ,  $SE = 0.05$ ,  $z = 4.13$ ,  $p < 0.001$ ), which can be seen in Figure 13.

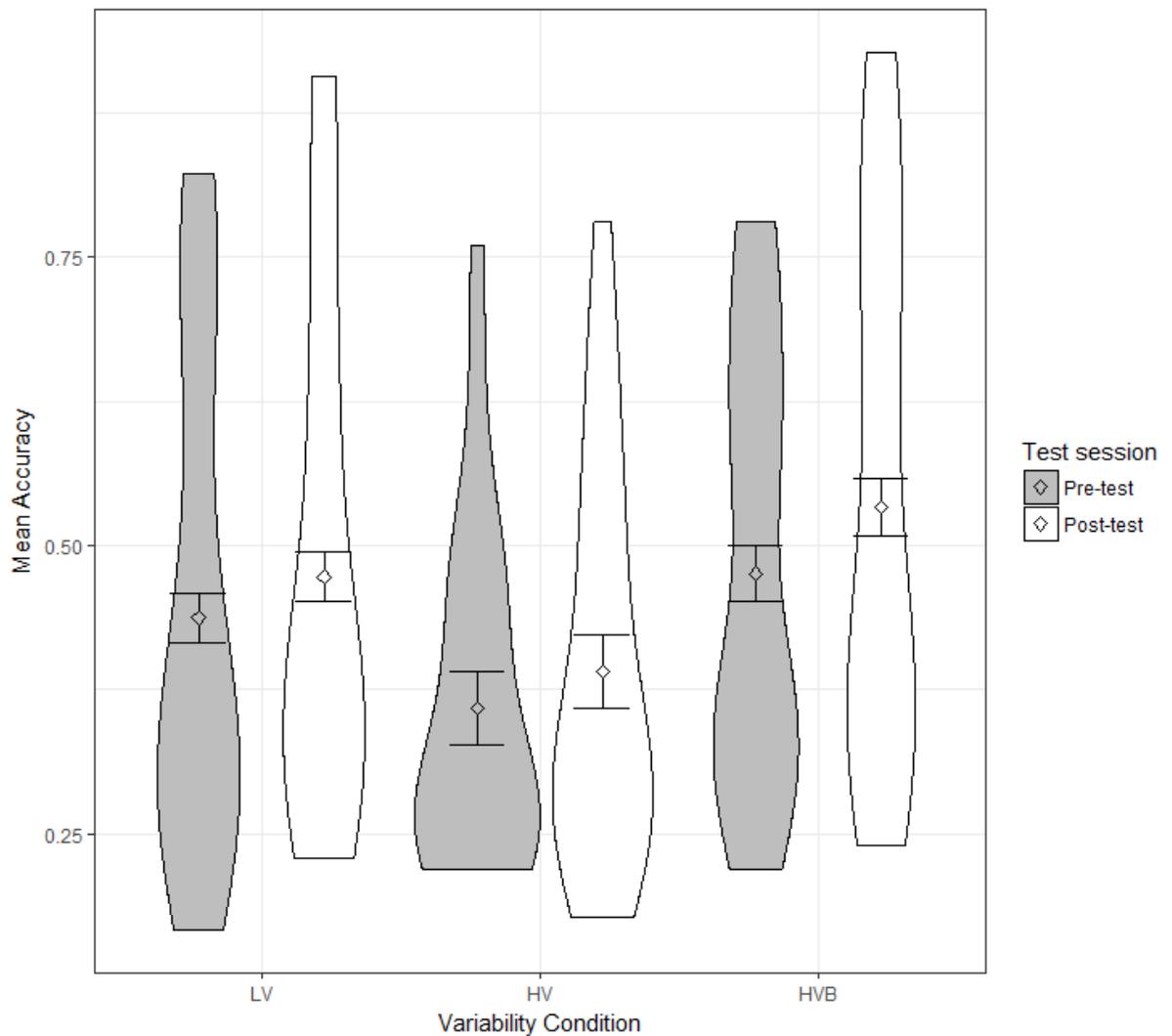


Figure 13 Mean proportion of correct for the LV (Low Variability), HV (High Variability) & HVB (High Variability Blocked) groups in Pitch Contour Perception Test. Error bars represents the 95% confidence intervals.

### 3.3.2.2 *Categorisation of Synthesised Tonal Continua*

Again, using the threshold provided by Sadakata and McQueen (2014) – i.e. removing data from participants with a slope measuring greater than 1.2, 10/20 participants in the current study failed the threshold. The fact that this was found in a second sample of participants provides further evidence that there might be a problem with the threshold given in the original Sadakata and McQueen (2014) paper. It is possible that this is due to a difference between participant groups. For example, all of their participants come from the Max Planck Institute for Psycholinguistics so they may be more motivated in these types of tasks. The same logistic

mixed effect model (Schultz, Llanos, & Francis, 2003, see section 2.3.2.2) was used to acquire the slope coefficients for each participant.

A correlation analysis was run between the two individual aptitude measures firstly looking just at the new HVB condition and then across all three groups. Results indicated there was no significant relationship between them for the HVB group only ( $r(18) = -0.11, p = 0.64$ ) but there was a significant positive relationship across all three groups ( $r(58) = .36, p < .01$ ). This, again, may suggest that the correlation found in Study 1 was due to coincidence.

A model was run to examine whether there was a difference between participants at pre-test. There was no improvement of this measure after training ( $\beta = -0.01, SE = 0.19, t = -0.03, p = 0.98$ ). There was a significant difference between LV and HVB contrast ( $\beta = 0.97, SE = 0.32, t = 3.01, p < .01$ ), suggesting that HVB participants had lower *Categorisation of Synthesized Tonal Continua* measure by accident at pre-test. However, there was no difference between HVB and HVB contrast which ( $\beta = 0.55, SE = 0.32, t = -1.29, p = 0.20$ ), can be seen in Figure 14.

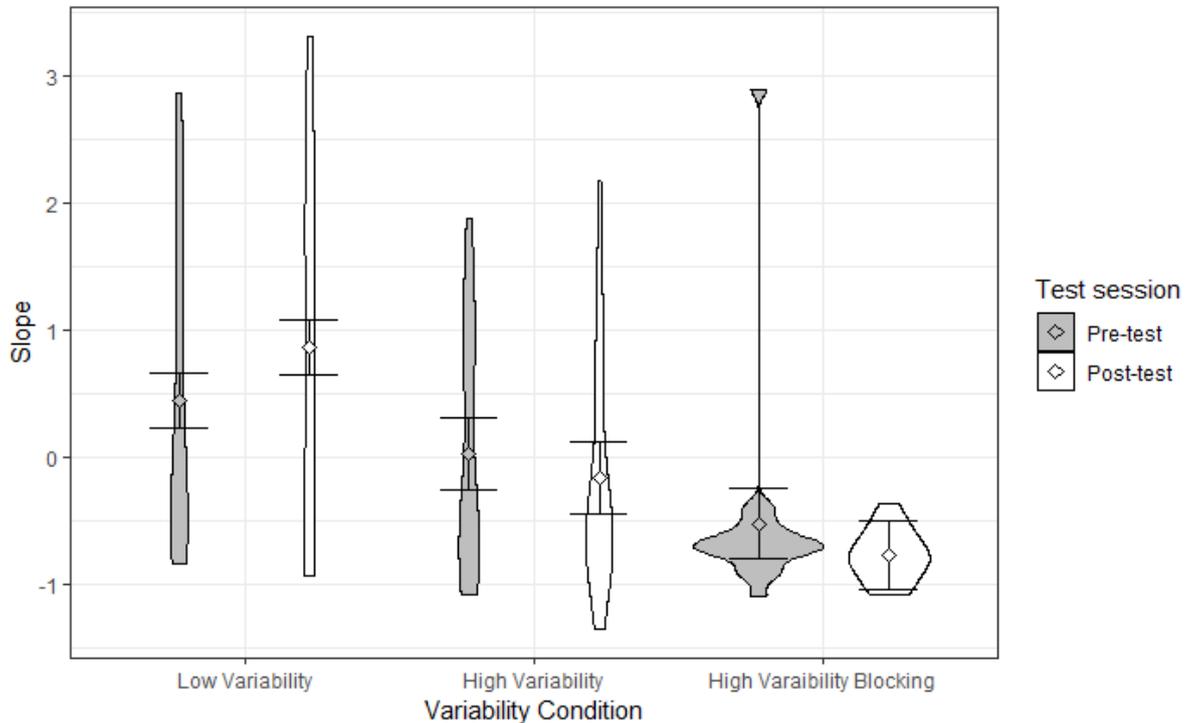


Figure 14 Slope measure for the LV (Low Variability), HV (High Variability) & HVB (High Variability Blocked) groups in the Categorisation of Synthesized Tonal Continua test. Error bars represents the 95% confidence intervals.

### 3.3.3 Training

A model containing data from all three conditions did not converge; two separate models, one including the LV and HVB conditions, and the other the HV 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 HVB; Model 2, HV versus HVB). The mean accuracy is displayed in Figure 15.

In both models, there was an effect of *training-session* (Model 1:  $\beta = 0.53$ ,  $SE = 0.04$ ,  $z = 12.17$ ,  $p < .001$ ; Model 2:  $\beta = 0.31$ ,  $SE = 0.04$ ,  $z = 8.50$ ,  $p < .001$ ): Participants' performance increased significantly over time, with additional training sessions. Overall, the LV group performed better than the HVB group ( $\beta = -0.83$ ,  $SE = 0.32$ ,  $z = -2.61$ ,  $p < .01$ ) and the HVB group outperformed the HV group ( $\beta = -0.73$ ,  $SE = 0.26$ ,  $z = -2.77$ ,  $p < .01$ ). The LV versus HVB contrast was also modulated by an interaction with *training-session* ( $\beta = -0.35$ ,  $SE = 0.08$

$z = -4.33, p < .001$ ). From Figure 14 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 few sessions. For the HV and the HVB group, there was no *participant-condition* x *training-session* interaction.

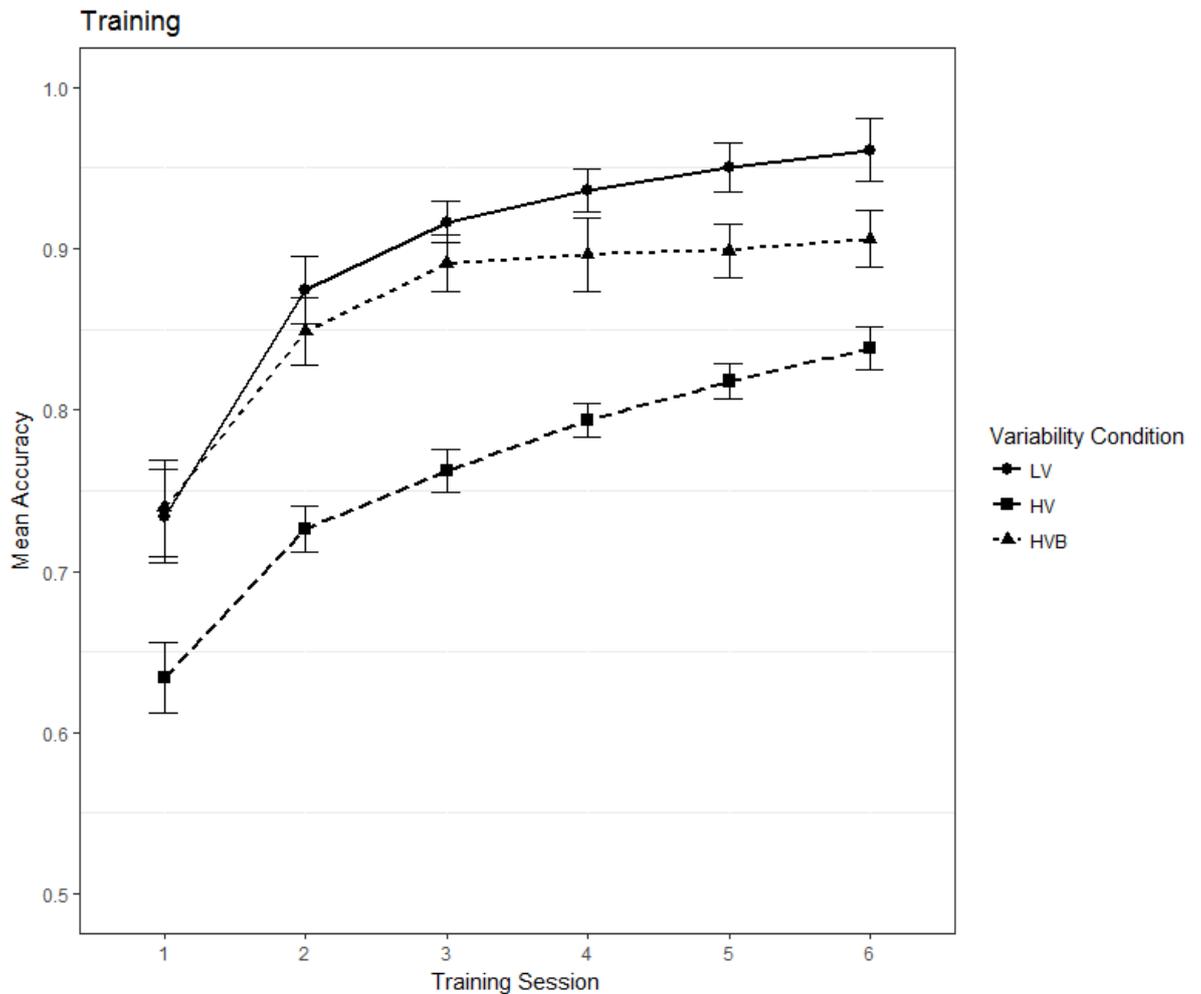


Figure 15 Mean proportion of correct in the Training task for the LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups in each session. Y-axis starts from chance level. Error bars show 95% confidence intervals.

### 3.3.4 Perceptual tests

#### 3.3.4.1 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 HVB,

HV versus HVB), *trial-type* (neutral versus easy, neutral versus hard) and *item-novelty* (trained item, untrained item). The mean accuracy is displayed in Figure 16.

At pre-test, there was no significant difference between the LV and the HVB groups ( $\beta = -0.12$ ,  $SE = 0.14$ ,  $z = -0.86$ ,  $p = .39$ ) nor between the HV and the HVB groups ( $\beta = -0.12$ ,  $SE = 0.14$ ,  $z = -0.97$ ,  $p = .39$ ), suggesting that the groups started at a similar level. However, performance with the “*untrained*” was significantly better than performance on the “*trained*” items at pre-test ( $\beta = 0.31$ ,  $SE = 0.06$ ,  $z = 4.95$ ,  $p < 0.001$ ), suggesting incidental differences between item sets. 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,  $\beta = 0.40$ ,  $SE = 0.08$ ,  $z = 5.09$ ,  $p < 0.01$ ) and “*neutral*” trials were marginally easier than “*hard*” trials (where one of the foil speakers had the odd gender out,  $\beta = -0.14$ ,  $SE = 0.08$ ,  $z = -1.81$ ,  $p = 0.07$ ).

Overall, participants’ performance increased significantly after training ( $M_{pre} = 0.59$ ,  $SD_{pre} = 0.21$ ,  $M_{post} = 0.66$ ,  $SD_{post} = 0.19$ ,  $\beta = 0.31$ ,  $SE = 0.05$ ,  $z = 6.54$ ,  $p < .001$ ). The interaction between *test-session* and *item-novelty* was not significant ( $\beta = 0.14$ ,  $SE = 0.09$ ,  $z = 1.49$ ,  $p = .14$ ), suggesting no evidence that training had a greater effect for trained words than for untrained words. Critically, there was no interaction with *test-session* for either the contrast between the LV versus the HVB conditions ( $\beta = -0.01$ ,  $SE = 0.12$ ,  $z = -0.11$ ,  $p = .91$ ) or the contrast between the HV versus the HVB conditions ( $\beta = -0.03$ ,  $SE = 0.12$ ,  $z = -0.23$ ,  $p = .82$ ) and they were not qualified by any higher level interactions with *item-novelty* (LV versus HVB:  $\beta = -0.13$ ,  $SE = 0.22$ ,  $z = -0.58$ ,  $p = 0.57$ ; HV versus HVB:  $\beta = -0.27$ ,  $SE = 0.23$ ,  $z = -1.21$ ,  $p = 0.23$ ). This suggests no evidence that the extent to which participants improved on this task between pre and post-test differed according to *variability-conditions*, or that this differed for *trained* versus *untrained* items.

Although not part of the key predictions, the analysis also examined 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 ( $\beta = -0.27$ ,  $SE = 0.11$ ,  $z = -2.39$ ,  $p = .02$ ) while the contrast between “neutral” and “hard” was not ( $\beta = 0.12$ ,  $SE = 0.11$ ,  $z = 1.06$ ,  $p = .29$ ). This was due to the fact that there was improvement for “neutral” ( $M_{pre} = 0.57$ ,  $SD_{pre} = 0.14$ ,  $M_{post} = 0.65$ ,  $SD_{post} = 0.15$ ) and “hard” trials ( $M_{pre} = 0.54$ ,  $SD_{pre} = 0.16$ ,  $M_{post} = 0.65$ ,  $SD_{post} = 0.15$ ) but not for “easy” trials ( $M_{pre} = 0.66$ ,  $SD_{pre} = 0.16$ ,  $M_{post} = 0.68$ ,  $SD_{post} = 0.15$ ).

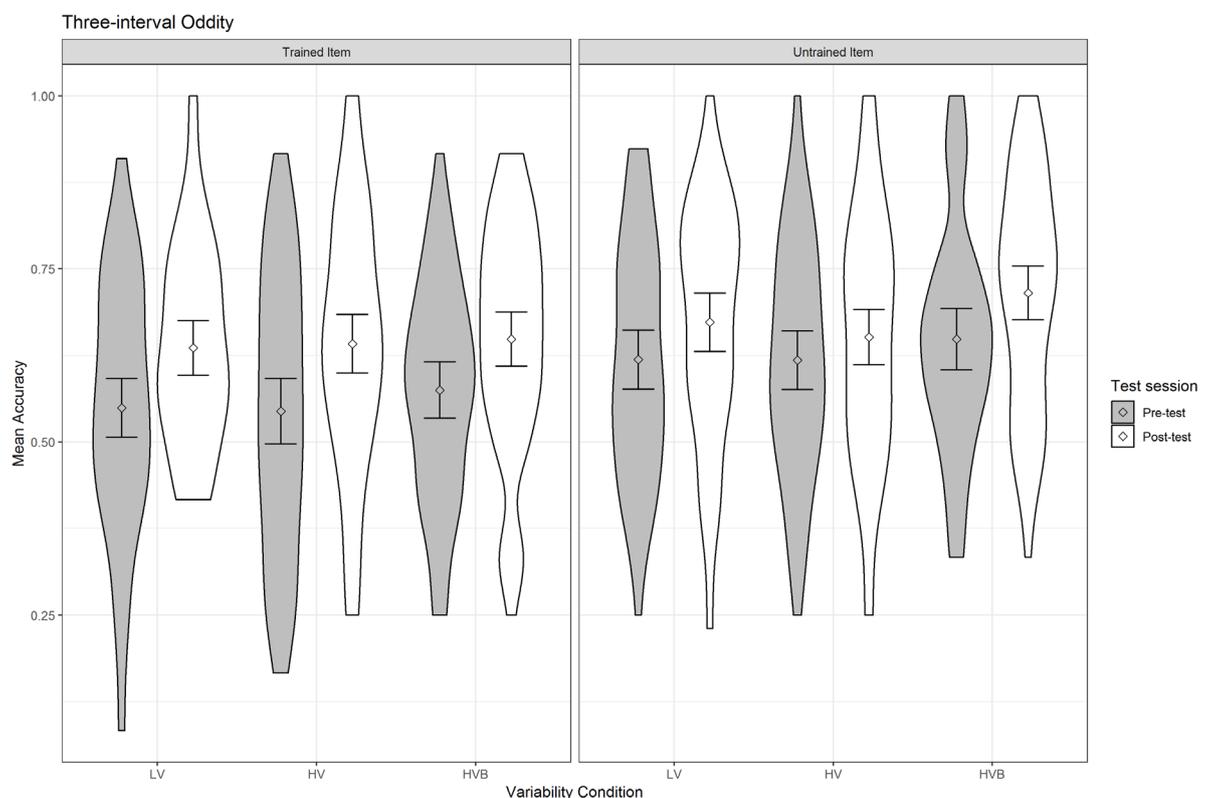


Figure 16 Mean proportion of correct in Three Interval Oddity task for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals.

### 3.3.4.2 Picture Identification

The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were the factor *voice-novelty* (*trained* voice, *untrained* voice) and the factor

*variability-condition* which had two contrasts (LV versus HVB, HV versus HVB). The mean accuracy is displayed in Figure 17.

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. There was no significant difference between the LV group and the HV group ( $\beta = 0.14$ ,  $SE = 0.32$ ,  $z = 0.44$ ,  $p = 0.66$ ) nor between the HV and the HVB group ( $\beta = -0.57$ ,  $SE = 0.31$ ,  $z = -1.81$ ,  $p = 0.07$ ). There was a significant interaction between *voice-novelty* and the LV-HVB contrast ( $\beta = 1.11$ ,  $SE = 0.36$ ,  $z = -3.08$ ,  $p < .01$ ). Breaking this down by *variability-condition*: For each condition there was significantly better performance with *trained* than *untrained* voices (LV:  $\beta = 1.83$ ,  $SE = 0.29$ ,  $z = 6.42$ ,  $p < 0.001$ ; HVB:  $\beta = 0.73$ ,  $SE = 0.26$ ,  $z = 2.82$ ,  $p < 0.01$ ), although it was larger in the LV condition. Breaking down by *voice-novelty*: For trained voice, performance was higher in the LV condition than in the HVB conditions, although the difference was not significant ( $\beta = -0.70$ ,  $SE = 0.45$ ,  $z = -1.55$ ,  $p = 0.12$ ). Importantly, for *untrained* voices, there was also no difference between the groups ( $\beta = -0.41$ ,  $SE = 0.27$ ,  $z = -1.51$ ,  $p = 0.13$ ), so that there was no evidence that the exposure to multi speaker input aided generalisation.

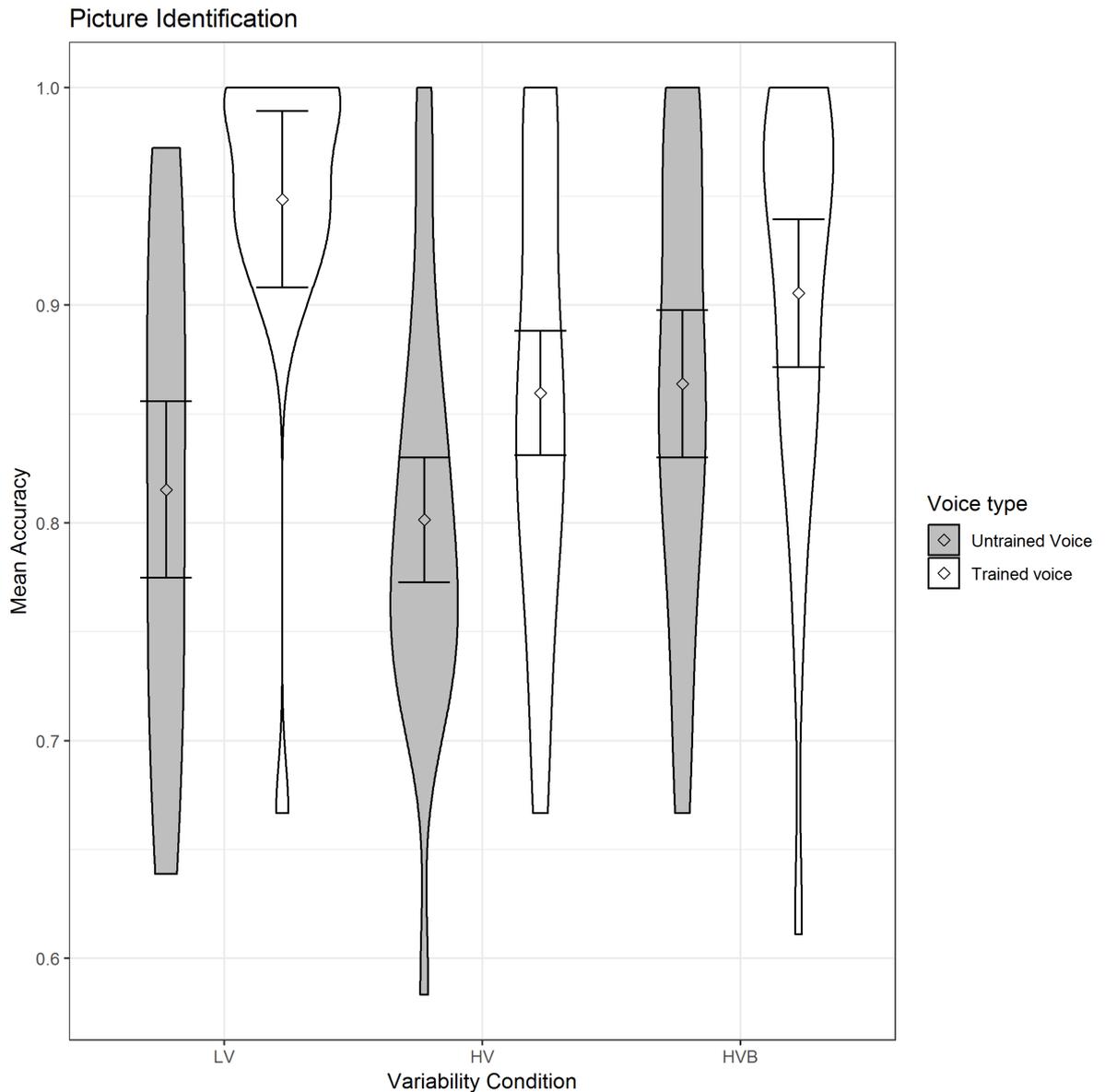


Figure 17 Mean proportion of correct of Picture Identification for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups for untrained voices and trained voices. Error bars show 95% confidence intervals.

### 3.3.5 Production tests

#### 3.3.5.1 Coding and inter-rater reliability analyses

The same methods were used for both production tests as described in section 2.3.5.1. Data were removed according to same criteria as before. In total, this resulted in 3.38% (359/10620) of production trials being removed from analysis (*Word Repetition*: Pre-test 1.98% (84/4248); Post-test 3.72% (158/4248); *Picture Naming* 5.51% (117/2124)). 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 in 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 all three 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  $kappa = 0.39$  (“fair agreement”), and for Pinyin accuracy  $kappa = 0.33$  (“fair agreement”; Landis & Koch, 1977). For the Picture Naming test, for Tone accuracy  $kappa = 0.67$  (“substantial agreement”) and for Pinyin accuracy  $kappa = 0.53$  (“moderate agreement”); For the Tone rating, the package “irr” in R was used to assess the intra-class correlation (McGraw & Wong, 1996) based on an average-measures, two-way mixed-effects model. For Word Repetition,  $ICC = 0.22$  and for Picture Identification  $ICC = 0.37$ ; according to Cicchetti (1994), values less than .40 are regarded as “poor”. Again, the inter-rater relationship for tone rating measure is still low, thus no analysis was performed. All of the analyses presented in Sections 3.3.5.2 and 3.3.5.3 were based on Rater 2 (the naive rater).

### 3.3.5.2 Word Repetition

#### 3.3.5.2.1 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 HVB, HV versus HVB) and *item-novelty* (trained, untrained). The mean accuracy, split by *test-session* and training condition, is shown in Figure 18.

At pre-test, there was no significant difference between the LV and the HVB group ( $\beta = 0.11, SE = 0.18, z = 0.64, p = 0.53$ ) nor between the HV and the HVB group ( $\beta = 0.01, SE = 0.18, z = 0.07, p = 0.94$ ), suggesting the groups started at a similar level. There was also no difference between *trained* and *untrained* items at pre-test ( $\beta = -0.02, SE = 0.07, z = 0.-0.26, p = 0.80$ ).

Across the three groups, participants' performance increased significantly after training ( $M_{pre} = 0.71, SD_{pre} = 0.09, M_{post} = 0.79, SD_{post} = 0.09, \beta = 0.40, SE = 0.08, z = 5.29, p < .001$ ). There was no significant difference in the improvement for *trained* and *untrained* items (*word-type* by *test-session* interaction:  $\beta = 0.13, SE = 0.10, z = 1.22, p = .22$ ). Critically, the interactions between the variability contrasts and *test-session* were not significant (LV versus HVB:  $\beta = 0.11, SE = 0.18, z = 0.62, p = 0.54$ ; HV versus HVB:  $\beta = 0.01, SE = 0.18, z = 0.07, p = 0.94$ ), and they were not qualified by any higher level interactions with *item-novelty* (LV versus HVB:  $\beta = 0.31, SE = 0.26, z = 1.21, p = 0.23$ ; HV versus HVB:  $\beta = 0.46, SE = 0.26, z = 1.80, p = 0.07$ ;). This suggests there is no evidence that participants' improvement in their production of tones was affected by their *variability-condition*, or that this differed for *trained* versus *untrained* items.

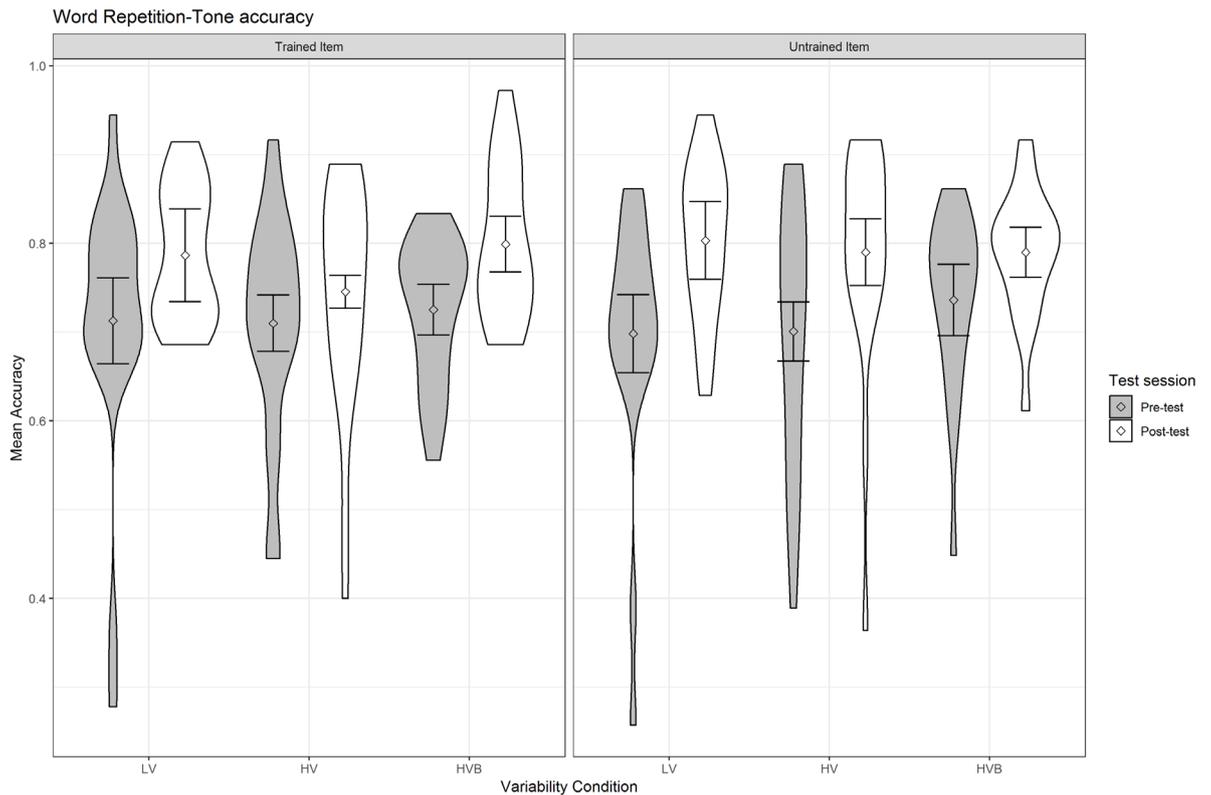


Figure 18 Accuracy of Word Repetition for LV (Low Variability), High Variability (HV) and High Variability Blocked (HVB) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals.

### 3.3.5.2.2 Pinyin accuracy

The predicted variable was whether the participants produced the correct string of phonemes (1/0) in each trial (as determined by Rater 2). The predictors were *test-session* (pre-test, post-test), *variability-condition* (LV versus HVB, HV versus HVB) and *item-novelty* (trained, untrained). Mean pinyin accuracy is displayed in Figure 19.

At pre-test, there was no significant difference between the LV and the HVB group ( $\beta = 0.03$ ,  $SE = 0.11$ ,  $z = 0.24$ ,  $p = 0.81$ ) nor between the HV and the HVB group ( $\beta = 0.01$ ,  $SE = 0.11$ ,  $z = 0.13$ ,  $p = .90$ ), suggesting that the groups started at a similar level. However, participants did better on untrained words than trained words at pre-test ( $\beta = 0.21$ ,  $SE = 0.07$ ,  $z = 3.11$ ,  $p < .01$ ), suggesting potential accidental differences in these items. Participants showed significant improvement after training ( $M_{pre} = 0.54$ ,  $SD_{pre} = 0.09$ ,  $M_{post} = 0.58$ ,  $SD_{post} = 0.19$ ,  $\beta = 0.15$ ,  $SE = 0.05$ ,  $z = 3.38$ ,  $p < .01$ ). However, there was no evidence that different

*variability-conditions* resulted in different amounts of improvement (*test-session* by LV versus HVB:  $\beta = 0.12$ ,  $SE = 0.11$ ,  $z = 1.08$ ,  $p = .28$ ; *test-session* by HV versus HVB:  $\beta = 0.17$ ,  $SE = 0.11$ ,  $z = 1.52$ ,  $p = 0.13$ ) or any interaction between *variability-condition*, *test-session* and *item-novelty* (LV versus HVB:  $\beta = 0.14$ ,  $SE = 0.22$ ,  $z = 0.64$ ,  $p = 0.52$ ; HV versus HVB:  $\beta = 0.25$ ,  $SE = 0.22$ ,  $z = 1.14$ ,  $p = 0.25$ ). This suggests there is no evidence that participants' improvement in pinyin accuracy was affected by their *variability-condition*, or that this differed for *trained* versus *untrained* items.

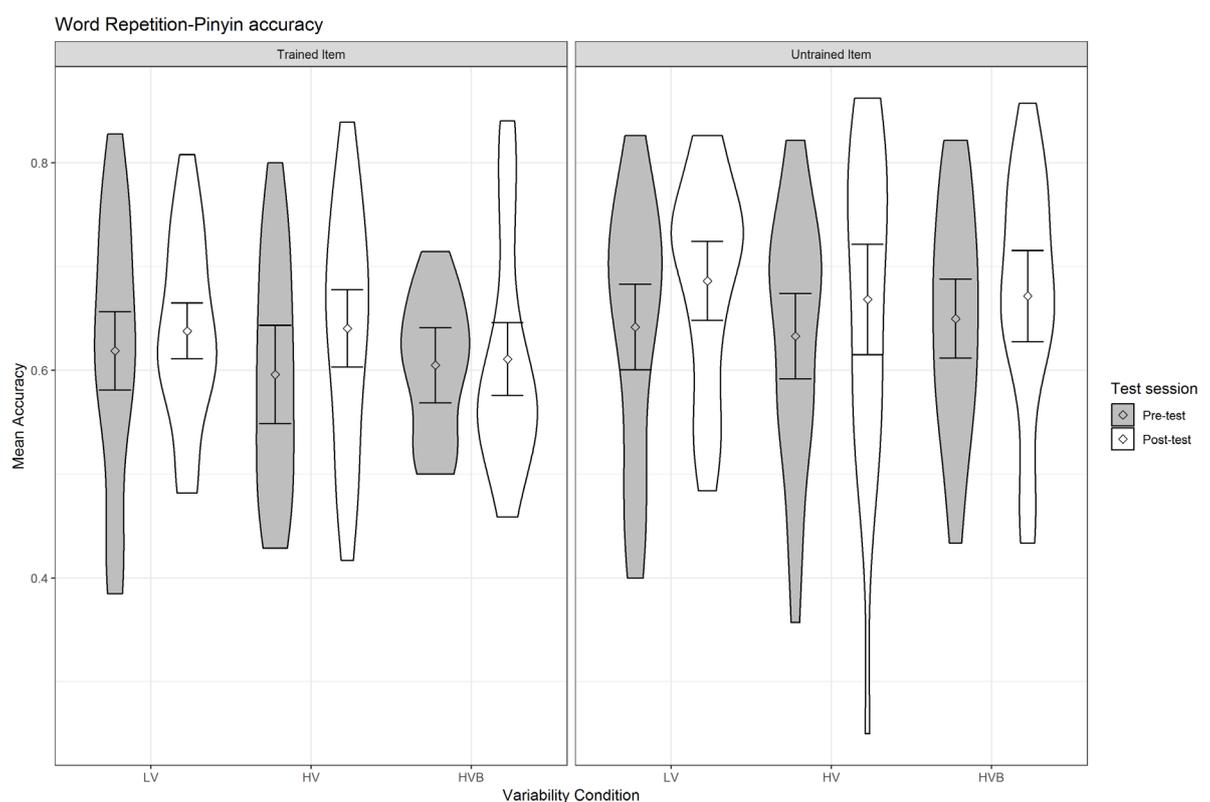


Figure 19 Mean pinyin accuracy of Word Repetition for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups in Pre- and Post-tests for trained and untrained items. Error bars show 95% confidence intervals.

### 3.3.5.3 Picture Naming

#### 3.3.5.3.1 Tone accuracy

The predicted variable was whether a correct response was given (1/0) on each trial (as identified by the coder). There was only one predictor, *variability-condition* (LV versus HVB, HV versus HVB) for both models. The descriptive statistics are displayed in Figure 20.

There was no significant difference between the LV-HVB contrast ( $\beta = 0.10$ ,  $SE = 0.19$ ,  $z = 0.52$ ,  $p = 0.61$ ) and between the HV-HVB contrast ( $\beta = -0.24$ ,  $SE = 0.19$ ,  $z = -1.26$ ,  $p = 0.21$ ). This suggests there is no evidence that participants' ability to produce the tones accurately differed according to their *variability-condition*.

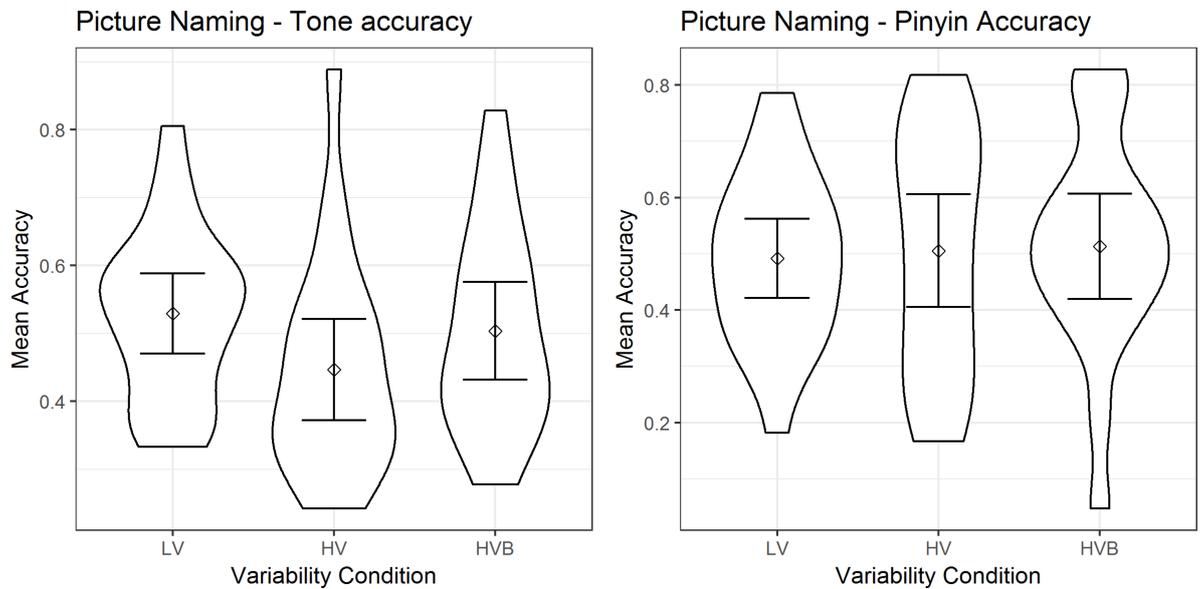


Figure 20 Tone accuracy and Pinyin accuracy of Picture Naming for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups. Error bars show 95% confidence intervals.

### 3.3.5.3.2 Pinyin Accuracy

The predicted variable was whether the participants produced the correct string of phonemes (1/0) in each trial and there was a single predictor *variability-condition* (LV versus HVB, HV versus HVB). For both models there was no significant difference between *variability-conditions* (LV versus HVB:  $\beta = -0.12$ ,  $SE = 0.23$ ,  $z = -0.51$ ,  $p = 0.61$ ; HV versus HVB:  $\beta = -0.03$ ,  $SE = 0.23$ ,  $z = -0.11$ ,  $p = 0.91$ ). This suggests there is no evidence that participants' pinyin accuracy differed according to their *variability-condition*.

### 3.3.6 Analyses with Individual Aptitude

Similar to Study 2, individual aptitudes were acquired by calculating the mean accuracy at pre-test on the *Pitch Contour Perception Test* and the slope coefficient at pre-test on the *Categorization of Synthesized Tonal Continua*. These scores were centered and used as a

continuous predictor (*aptitude*) and added to each of the models reported above, and corresponding correlations relating to the main hypotheses were also included (see Table 10 and Table 11). Specifically for this study, it was expected that the HVB group showed better performance than the HV group. Based on Perrachione et al. (2011), Sadakata and McQueen (2014) and the results from Study 1, HV should benefit high aptitude participants only, while low variability would benefit low aptitude participants only.

The results with *Pitch Contour Perception Test* as the ID measure are shown in Table 10. *Aptitude* is a positive predictor of performance in each of the tests and in training, with *p*-values significant or marginal in each case. 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 the 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 Pitch Contour Perception Test, for the purposes of visualisation, Figure 21 (Three Interval Oddity task and Training task), Figure 22 (Picture Naming and Picture Identification) and Figure 23 (Word Repetition) use the mean accuracy for participants split into aptitude groups using a median split based on their Pitch Contour Perception Test score. In sum, participants with higher aptitude measures 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.

The results of *Categorisation of Synthesized Tonal Continua* as the ID measure are shown in Table 11. The only significant result was on Word Repetition Tone accuracy measure where an *aptitude* x *test-session* x *LV-HVB* contrast interaction was revealed. However, post-hoc analysis revealed that there was a *negative* relationship between *aptitude* and participants' learning outcome for the HVB group ( $\beta = -0.25$ ,  $SE = 0.12$ ,  $z = -0.23$ ,  $p = 0.03$ ) although not

for the LV group ( $\beta = 0.05$ ,  $SE = 0.15$ ,  $z = 0.35$ ,  $p = 0.73$ ). This negative relationship can be seen in Figure 24. However it is not what is predicted by the hypothesis.

*Table 10* Statistics analysis with Pitch Contour Perception Test as the individual difference measure.

<b>Data Set</b>	<b>Coefficient Name</b>	<b>Statistics</b>
<i>Word</i>	<b>Aptitude</b>	<b><math>\beta = 0.07</math>, <math>SE = 0.03</math>, <math>z = 2.35</math>, <math>p = .019</math></b>
<i>Repetition:</i>	Aptitude by <i>Test-Session</i>	$\beta = 0.03$ , $SE = 0.04$ , $z = 0.72$ , $p = .473$
<i>Tone Accuracy</i>	Aptitude by LV-HVB Contrast by <i>Test-</i>	$\beta = -0.13$ , $SE = 0.10$ , $z = -1.35$ , $p = .176$
<i>(Pre/Post)</i>	<i>Session</i>	
	Aptitude by HV-HVB Contrast by <i>Test-</i>	$\beta = -0.08$ , $SE = 0.12$ , $z = -0.66$ , $p = .507$
	<i>Session</i>	
	Aptitude by LV-HVB Contrast by <i>Test-</i>	$\beta = -0.07$ , $SE = 0.13$ , $z = -0.50$ , $p = .61$
	<i>Session</i> by <i>Item-Novelty</i>	
	Aptitude by HV-HVB Contrast by <i>Test-</i>	$\beta = -0.21$ , $SE = 0.17$ , $z = -1.28$ , $p = .202$
	<i>Session</i> by <i>Item-Novelty</i>	
<i>Three Interval</i>	<b>Aptitude</b>	<b><math>\beta = 0.07</math>, <math>SE = 0.03</math>, <math>z = 2.19</math>, <math>p = .029</math></b>
<i>Oddity</i>	Aptitude by <i>Test-Session</i>	$\beta = 0.01$ , $SE = 0.03$ , $z = 0.31$ , $p = .757$
<i>(Pre/Post)</i>	Aptitude by LV-HVB Contrast by <i>Test-</i>	$\beta = -0.05$ , $SE = 0.06$ , $z = -0.83$ , $p = .410$
	<i>Session</i>	
	Aptitude by HV-HVB Contrast by <i>Test-</i>	$\beta = 0.003$ , $SE = 0.07$ , $z = 0.05$ , $p = .958$
	<i>Session</i>	
	Aptitude by LV-HVB Contrast by <i>Test-</i>	$\beta = -0.06$ , $SE = 0.11$ , $z = -0.52$ , $p = .604$
	<i>Session</i> by <i>Item-Novelty</i>	
	Aptitude by HV-HVB Contrast by <i>Test-</i>	$\beta = -0.18$ , $SE = 0.14$ , $z = -1.32$ , $p = .187$
	<i>Session</i> by <i>Item-Novelty</i>	

<i>Training</i>	<b>Aptitude</b>	<b><math>\beta = 0.13</math>, SE = 0.048, z = 2.70, p = .007</b>
	Aptitude by LV-HVB Contrast	$\beta = -0.03$ , SE = 0.10, z = -0.26, p = 0.796
	Aptitude by HV-HVB Contrast	$\beta = -0.06$ , SE = 0.12, z = -0.53, p = .596
<i>Picture</i>	<b>Aptitude</b>	<b><math>\beta = 1.48</math>, SE = 0.08, z = 1.96, p = .050</b>
<i>Identification</i>	Aptitude by Voice Novelty	$\beta = -0.03$ , SE = 0.07, z = -0.33, p = .745
<i>(Post Only)</i>	Aptitude by LV-HVB Contrast	$\beta = -0.01$ , SE = 0.17, z = -0.09, p = .932
	Aptitude by HV-HVB Contrast	$\beta = -0.04$ , SE = 0.19, z = -0.19, p = .847
	Aptitude by LV-HVB Contrast by <i>Voice-Novelty</i>	$\beta = 0.11$ , SE = 0.19, z = 0.57, p = .566
	Aptitude by HV-HVB Contrast by <i>Voice-Novelty</i>	$\beta = 0.45$ , SE = 0.21, z = 2.15, p = .031
<i>Picture</i>	<b>Aptitude</b>	<b><math>\beta = 0.08</math>, SE = 0.04, z = 1.89, p = 0.059</b>
<i>Naming: Tone</i>	Aptitude by LV-HVB Contrast	$\beta = -0.12$ , SE = 0.10, z = -1.22, p = .224
<i>Accuracy</i>	Aptitude by HV-HVB Contrast	$\beta = -0.21$ , SE = 0.11, z = -1.80, p = .073

*Table 11* Statistics analysis with Categorisation of Synthetized Tonal Continua as the individual difference measure.

<b>Data Set</b>	<b>Coefficient Name</b>	<b>Statistics</b>
	Aptitude	$\beta = 0.01$ , SE = 0.06, z = 0.22, p = .829
	Aptitude by <i>Test-Session</i>	$\beta = -0.11$ , SE = 0.08, z = -1.51, p = .132

<i>Word</i>	<b>Aptitude by LV-HVB Contrast by Test-</b>	<b><math>\beta = 0.36</math>, SE = 0.18, z = 2.02, p = .043</b>
<i>Repetition:</i>	<b><i>Session</i></b>	
<i>Tone Accuracy</i>	Aptitude by HV-HVB Contrast by Test-	$\beta = 0.12$ , SE = 0.19, z = 0.63, p = .530
<i>(Pre/Post)</i>	<i>Session</i>	
	Aptitude by LV-HVB Contrast by Test-	$\beta = -0.27$ , SE = 0.25, z = -1.07, p = .286
	<i>Session by Item-Novelty</i>	
	Aptitude by HV-HVB Contrast by Test-	$\beta = -0.32$ , SE = 0.29, z = -1.10, p = .272
	<i>Session by Item-Novelty</i>	
<hr/> <i>Three Interval</i>	Aptitude	$\beta = 0.08$ , SE = 0.06, z = 1.38, p = .167
<i>Oddity</i>	Aptitude by Test-Session	$\beta = -0.02$ , SE = 0.05, z = -0.49, p = .626
<i>(Pre/Post)</i>	Aptitude by LV-HVB Contrast by Test-	$\beta = 0.0001$ , SE = 0.12, z = 0.001, p
	<i>Session</i>	= .999
	Aptitude by HV-HVB Contrast by Test-	$\beta = -0.002$ , SE = 0.13, z = -0.02, p
	<i>Session</i>	= .987
	Aptitude by LV-HVB Contrast by Test-	$\beta = 0.35$ , SE = 0.22, z = 1.57, p = .115
	<i>Session by Item-Novelty</i>	
	Aptitude by HV-HVB Contrast by Test-	$\beta = 0.18$ , SE = 0.26, z = 0.68, p = .497
	<i>Session by Item-Novelty</i>	
<hr/> <i>Training</i>	Aptitude	$\beta = 0.12$ , SE = 0.09, z = 1.38, p = .174
	Aptitude by LV-HVB Contrast	$\beta = 0.25$ , SE = 0.18, z = 1.39, p = .166
	Aptitude by HV-HVB Contrast	$\beta = 0.03$ , SE = 0.19, z = 0.14, p = .892
<hr/> <i>Picture</i>	Aptitude	$\beta = 0.09$ , SE = 0.13, z = 0.66, p = .507
<i>Identification</i>	Aptitude by Voice Novelty	$\beta = -0.12$ , SE = 0.14, z = -0.81, p = .420
<i>(Post Only)</i>	Aptitude by LV-HVB Contrast	$\beta = 0.40$ , SE = 0.33, z = 1.24, p = .214
	Aptitude by HV-HVB Contrast	$\beta = 0.12$ , SE = 0.34, z = 0.36, p = .722

Aptitude by LV-HVB Contrast by *Voice-*  $\beta = 0.24$ ,  $SE = 0.36$ ,  $z = 0.65$ ,  $p = .514$

*Novelty*

Aptitude by HV-HVB Contrast by *Voice-*  $\beta = 0.27$ ,  $SE = 0.35$ ,  $z = 0.75$ ,  $p = .452$

*Novelty*

---

<i>Picture</i>	Aptitude	$\beta = -0.01$ , $SE = 0.08$ , $z = -0.13$ , $p =$
		0.894
<i>Naming: Tone</i>		
<i>Accuracy</i>	Aptitude by LV-HVB Contrast	$\beta = 0.19$ , $SE = 0.20$ , $z = 0.96$ , $p = .336$
	Aptitude by LV-HV Contrast	$\beta = -0.04$ , $SE = 0.22$ , $z = -0.17$ , $p = .869$

---

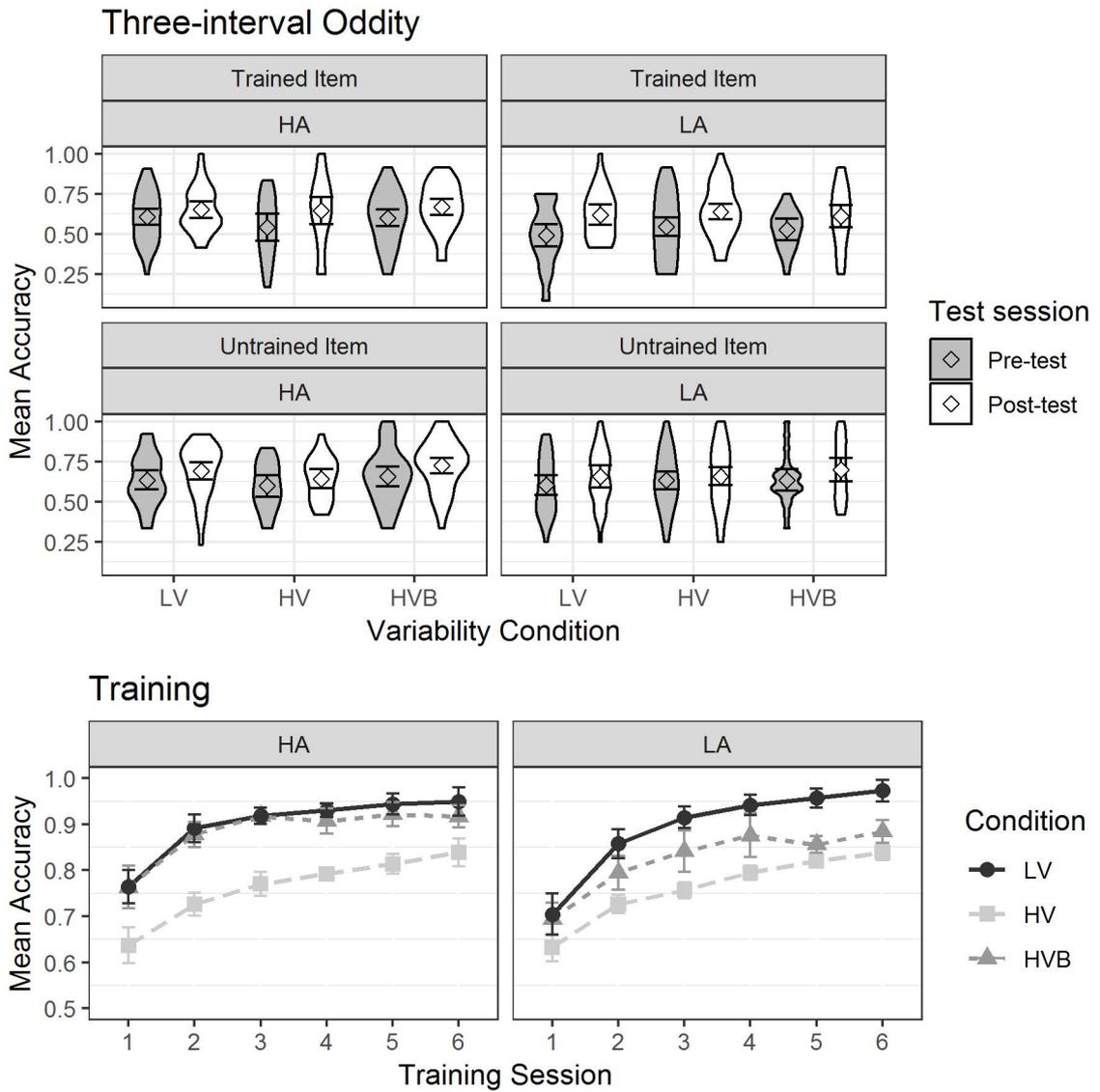


Figure 21 Accuracy in the Three Interval Oddity and Training data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, split by high versus low aptitude in the Pitch Contour Perception Test task. Error bars show 95% confidence interval.

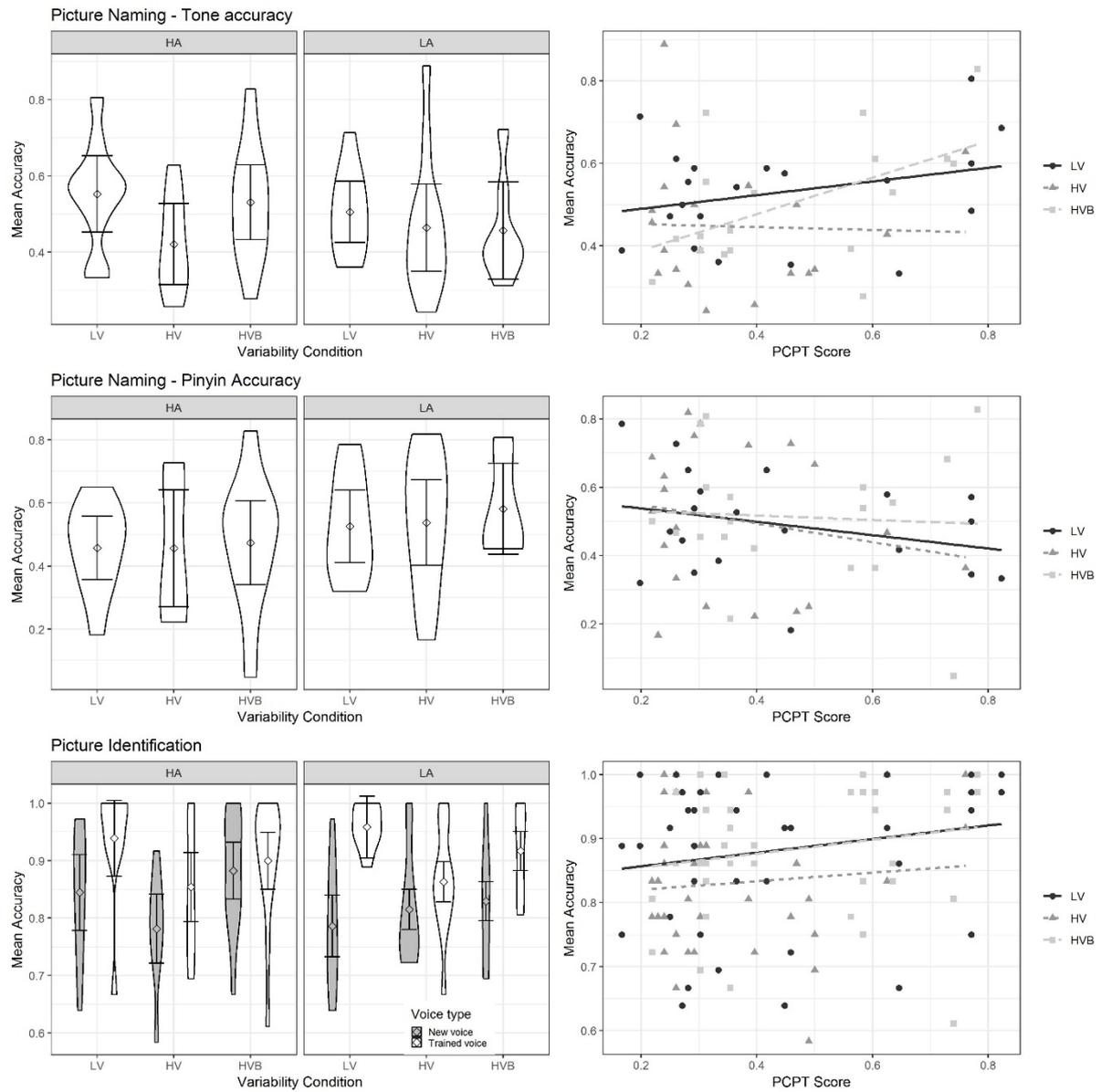


Figure 22 Accuracy in the Picture Naming and Picture Identification data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, split by high versus low aptitude in the Pitch Contour Perception Test. Error bars show 95% confidence intervals.

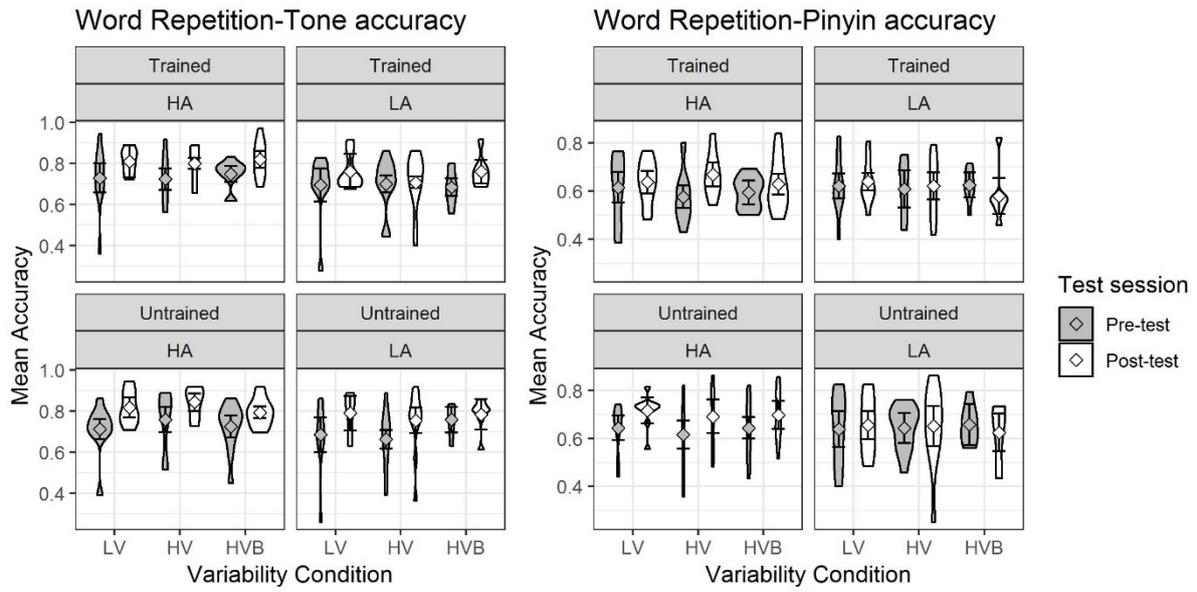


Figure 23 Accuracy in the Word Repetition data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, split by high versus low aptitude in the Pitch Contour Perception Test task. Error bars show 95% confidence interval

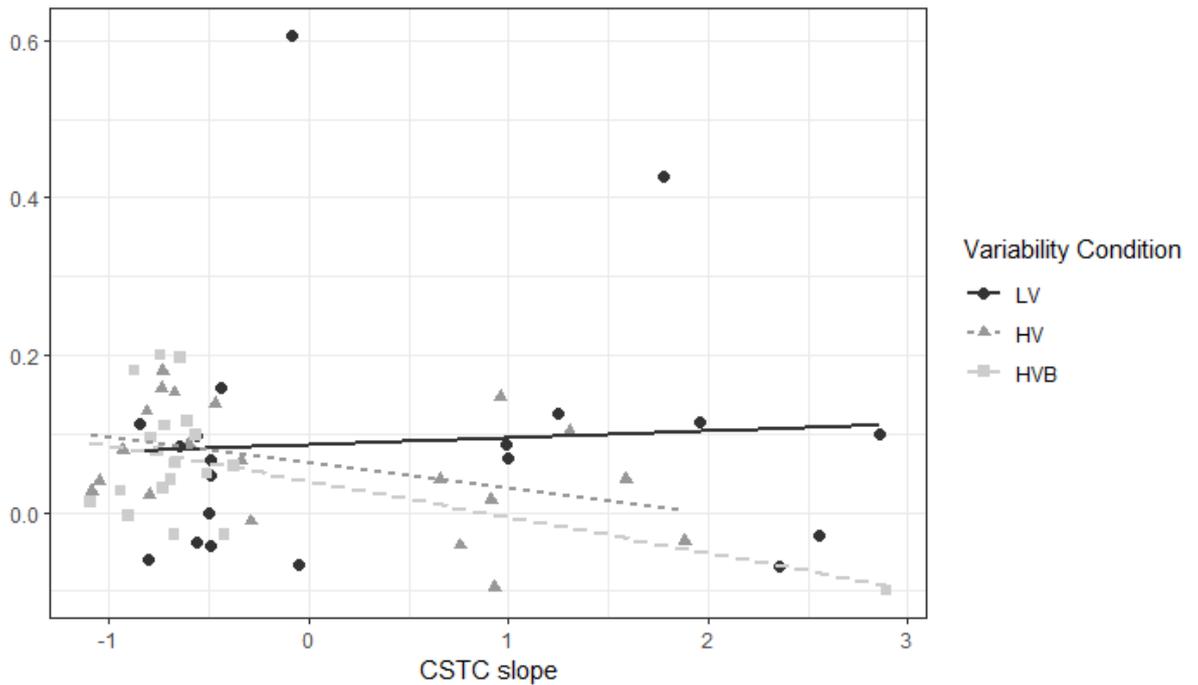


Figure 24 Word Repetition tone accuracy data for LV (Low Variability), HV (High Variability) and HVB (High Variability Blocked) training groups, with x-axis as the aptitude in the Categorisation of Synthesized Tonal Continua, and y-axis as the improvement from pre- to post-training.

### 3.3.7 Bayes Factor Analyses

In both the analyses reported in Chapter 2 and the analyses reported above, there was no evidence – in any of the tests – for neither of the two key hypotheses: (1) the hypothesis that training with multiple speakers leads to greater generalisation to new speakers than training with a single speaker *or* (2) the hypothesis that there is an interaction between the variability of the training materials and participant aptitude, such that higher aptitude participants benefit more from training with multiple speakers while lower aptitude participants benefit more from training with a single speaker. However, as noted above a non-significant result ( $p > .05$ ) does *not* tell whether there is enough evidence for the null, as opposed to no evidence for any conclusion at all, or even evidence against the null so these analyses should not reduce the confidence in either of the hypotheses.. In this section I report supplemental analyses using Bayes factors to evaluate the null results found in Study 1 and Study 2.

#### 3.3.7.1 H1: Greater generalisation - to novel voices and in production - in the multiple speaker conditions (HV and HVB) than in the low variability condition (LV)

The aim was to compute Bayes Factors comparing this hypothesis to the null for each of the data sets. To have maximum evidence, the model pooled the HV and HVB conditions and contrast this with the LV condition. For the post-tests the interest was in the evidence for a main effect of this contrast. For the pre-to-post tests, the main interest was on the interaction between this contrast and session. To further maximize evidence, for the Three Interval Oddity test and Word Repetition tests the model combined the trained and untrained items (since *both* types of item involve generalisation to an untrained voice and thus should benefit from high variability training), however in the Picture Identification test the model did excluded trained *voice* test items, since the benefit of high variability training was not predicted for these items. For the production measures, the major aim was to explore whether there was a HV benefit for the tone learning measure and the Pinyin measure (the latter given that Barcroft and Sommers (2014) found a benefit of multi-speaker training in their vocabulary recall task).

Bayes factors were computed following Dienes (2014) and Dienes, Coulton and Heather (2018). To compute a Bayes factor ( $B$ ) it is necessary to have both a model of the data and a model of H1. The model of the data is an estimate of the mean difference for the contrast in question, and of the standard error. Here, these estimates were acquired by running logistic mixed models and taking the betas and standard errors for the relevant coefficients (note that this allows the model to meet normality assumptions by continuing to work within log-odds space). The models ran here were similar to the previous analyses but with variability-condition coded as a centered contrast between LV and the HV+HVB conditions, and other factors combined/excluded as described in the previous paragraphs.

H1 was modelled using a half-normal distribution with a mode of 0 and a standard deviation  $x$  which is set to be a rough estimate of the predicted difference for this contrast. This allows for possible effects between 0 and twice the predicted effect, with values closer to 0 being more likely (Dienes, 2014).

In the absence of any prior data using sufficiently similar materials, and since it was less meaningful to use unprincipled default values, the  $x$  estimation for each contrast was achieved by using the scale and/or values from elsewhere in the data (see Dienes 2014, 2015 for a related approach). Specifically, for each of the cases where it predicted a main effect (Picture Identification and Picture Naming),  $x$  was set as the difference between the grand mean (the Intercept - since it used a centered coding) and an estimate of minimal possible performance on the task. The logic is as follows: The *maximum* difference between conditions is seen if *low variability* participants show baseline performance and *high variability* participants show performance greater than baseline. In this case, if performance on this test is  $p$  (so the grand mean is  $\bar{p}$ ) and the baseline is  $b$ , the difference in  $p$  between the two conditions will be equal to:  $2(\bar{p}-b)$ . This gives an estimate of the *maximum* value of  $x$ ; since the model is using a half normal distribution with a mean of zero, the maximum value should be equal to

approximately  $2SD$ , and the estimate  $x$  of the standard deviation can be set to be equal to *half* of this value (i.e.  $x = \bar{p}-b$ ). Baseline performance depends on the task: for the 2AFC Picture Identification task it is chance (50% = 0 in log odds space); for the Picture Naming, tone measure, I assume a  $\frac{1}{4}$  chance of identifying the correct one (25% = -1.099 in log odds space); for Picture Naming, Pinyin measure, there is no absolute chance level thus the model took the minimal performance as making one correct response in the test<sup>5</sup> (i.e.  $1/72 = -4.263$  in log odds space). For the cases where it is estimating an interaction between *test-session* and *variability-condition*  $x$  is set as equal to the mean increase in performance from *pre-* and *post-* test across conditions (main effect of *test-session*). The logic is as follows: the *maximum* difference is seen if *low variability* participants show no effect of *test-session* (no improvement) and *high variability* participants show a positive effect of *test-session*. In this case, if the mean effect of *test-session* is  $\bar{t}$ , the difference in  $t$  between the two conditions will be equal to  $2\bar{t}$ . Again, we can set our estimate of  $x$  to be half this value (i.e.  $x = \bar{t}$ ).

The interpretation of BFs used the following conventions:  $B < 1/3$  indicates substantial evidence for the null,  $B > 3$  indicates substantial evidence for H1, values between  $1/3$  and  $3$  indicate that the data collected do not sensitively distinguish H0 from H1 (Jeffreys 1961; Dienes 2008). Since there is subjectivity in how the values for H1 are determined, the results indicate the robustness of Bayesian conclusions by reporting a robustness region<sup>6</sup> for each  $B$ , which gives the range of values of the scale factor  $x$  that qualitatively support the same

---

<sup>5</sup> Note that it is impossible to compute log-odds of 0.

<sup>6</sup> To find out about the range of values, in each case I started at 0 (i.e. no difference between conditions) and went through 100 equal steps up to a value *max*; *max* was what I considered to be the largest possible difference between the two conditions in log odds space given the scale. Since all of the outcome measures are binary, I set the value of *max* to 5, equivalent to 99% accuracy (recall that log odds for 100% cannot be computed). In some cases I did not find the end of the robustness region within this range, in this case: I denoted the end of the range as “,> *max*”. In addition, for  $BF < 1/3$ , the end of the robustness regions is always infinity, as written in  $\infty$ .

conclusion (i.e. evidence as supporting H0, or as supporting H1, or there not being much evidence at all). Note that for evidence for H0, the maximum  $x$  is always infinity. The results are reported in Table 12. It can be seen there is substantial or strong evidence for the null for every test except for the Word Repetition test for the Pinyin accuracy measure, where the evidence is ambiguous, and that the robustness regions indicate that it would continue to have evidence for the null even with smaller estimates of the scale factor  $x$ .

*Table 12* Bayesian analysis for Picture Identification, Picture Naming, Word Repetition and Three Interval Oddity, with red cells representing evidence for the Null and yellow cells representing ambiguous results.

<b>Contrast</b>	<b>Mean difference</b>	<b>Stand. Error</b>	<b>H1 estimate <math>x</math></b>	<b>Bayes Factor (<math>B</math>)</b>	<b>Robustness Region</b>
Picture ID (Novel voice only) HV+ HVB > LV	0.13	0.228	1.71	0.219	1.11 : $\infty$
Picture Naming, (Tone accuracy) HV+ HVB > LV	-0.225	0.168	1.076	0.067	0.202 : $\infty$
Picture Naming (Pinyin Accuracy) HV+ HVB > LV	0.104	0.196	4.05	0.08	0.101 : $\infty$
Word Repetition (Tone accuracy) <i>test-session</i> by HV+ HVB > LV	-0.108	0.157	0.395	0.239	0.303 : $\infty$
Word Repetition (Pinyin accuracy) <i>test-session</i> by HV+ HVB > LV	0.095	-0.034	0.152	0.421	0 : 0.202
Three Interval Oddity <i>test-session</i> by HV+ HVB > LV	-0.001	0.1	0.31	0.303	0.303 : $\infty$

3.3.7.2 *H1: There is an interaction between an individual's tone-aptitude and variability-condition, such that participants with greater tone-aptitude show greater performance following the multiple speaker conditions (HV and HVB) and those with lesser tone aptitude show greater performance in the single speaker condition (LV)*

For this hypothesis, the same approach was taken as above except that it also included Bayes Factors for Training data, and for the Picture Identification test, both trained voice and untrained voice data were combined – pooling the two in order to maximize available evidence. This is because this interaction has been reported with trained items (Sadakata & McQueen, 2013) as well as untrained items (Perrachione et al., 2011). Again, the model also combined the HV and HVB conditions except for training where the analysis was run separately at the LV versus HV and LV versus HVB contrasts, since previous analyses from Study 1 and Study 2 has demonstrated that HV and HVB were different (HVB participants show higher performance). The model also combined the evidence from trained and untrained *items* in the pre- to post- tests. For the post-session only tests, the main focus was on the evidence for an interaction between the *variability-condition* contrast and *aptitude*. For the tests which appeared both pre- and post- training, the analysis focused on the interaction between the *variability-condition* contrast, *aptitude* and *test-session*. For training the current analyse also looked at the evidence for an interaction between each *variability-condition* contrast and *aptitude* (a more complex model containing the interaction with training-session did not converge). As in the frequentist analyses of aptitude, for the production measures – Word Repetition and Picture Naming – *no* analyses was performed on the pinyin measures since the *aptitude* measure is relevant only to tone learning.

Bayes Factors were computed following the same procedure as in Section 3.3.7.1 and again derived the estimates of the scale factor  $x$  - the difference predicted under H1 - using the scale and/or values from elsewhere in the data. Specifically, for each of the cases where it predicted a two-way interaction between *variability-condition* and *aptitude*,  $x$  was set as equal to the mean effect of *aptitude* across conditions (main effect of *aptitude*)<sup>7</sup>. The logic is as

---

<sup>7</sup> An alternative which would be more equivalent to the other BF analyses would be to inform the effect using the value of the two-way interaction of aptitude: test-session. I did not do this since I did not find an effect of this two-way interaction in either data set.

follows: The *maximum* difference is seen if *low variability* participants show no effect of *aptitude* and the *high variability* participants show a positive effect of *aptitude* (note that a negative effect of *aptitude* is not expected in any condition). In this case, if the mean effect of *aptitude* is  $\bar{a}$ , the difference in *a* between the two conditions will be equal to  $2\bar{a}$ . Again, the estimate of *x* – the *SD* of the half normal – was set to be half this maximum value i.e.  $x = \bar{a}$ . For the cases where the three-way interaction between *aptitude*, *test-condition* and *test-session* was involved, the estimate was based on half the difference between the maximal effect of *aptitude* (*maxA* – taken from the scale) and their actual *aptitude* score at pre-test (*baselineA* – taken from the data). The logic is as follows: The maximal effect of the interaction would be seen if participants in the *low variability* condition showed the same baseline effect of *aptitude* at *pre-test* and at *post-test* (*bA*), whereas participants in the *high variability* condition showed maximal improvement at post-test (*maxA*). In this case, the interaction between *aptitude* and session for the *high variability* group would be equal to: *maxa* – *ba*. Again, the estimate of *x* – the *SD* of the half normal – can be set to be half this maximum value, i.e.  $x = \frac{\text{maxa} - \text{ba}}{2}$ .

The maximum effect of *aptitude* was computed from the scale and the length of the *aptitude* predictor. Specifically, the assumption was that the maximal effect of *aptitude* would be obtained if participants with maximal *aptitude* were at ceiling (71/72 correct – log odds 4.263) and those with minimal *aptitude* were at chance (25% in Word Repetition, Tone Accuracy, log odds= 1.099; 33.33% in Three Interval Oddity, log odds = 0.693). This range was then further divided by the length of the *aptitude* predictor to obtain a measure of a one-step change in *aptitude*.

The results for Pitch Contour Perception Test as the *aptitude* measure are summarised in Table 13. It can be seen that although there is more evidence for the null than H1 in each case (i.e.  $BF < 1$ ) there is *no* substantial evidence for the null over H1 in any case. Thus, it is impossible to draw any inferences about the interaction from this data. Note that, in most cases,

the robustness regions indicate that even if the scale factor  $x$  was twice as large, i.e. corresponding to the *maximum* value I expected, the  $B$  would be ambiguous.

*Table 13* Bayesian analysis with PCPT as the ID measure, with red cells representing evidence for the Null and yellow cells representing ambiguous results.

<b>Contrast</b>	<b>Mean difference</b>	<b>Stand. Error</b>	<b>H1 estimate <math>x</math></b>	<b>Bayes Factor (<math>B</math>)</b>	<b>Robustness Region</b>
Picture ID, (Tone accuracy) <i>aptitude</i> by HV+ HVB > LV	0.006	0.127	0.171	0.617	0: 0.354
Picture Naming, (Tone accuracy) <i>aptitude</i> by HV+ HVB > LV	0.042	0.083	0.099	0.904	0: 0.354
Three Interval Oddity (Tone accuracy) <i>aptitude</i> by <i>test-session</i> by HV+ HVB > LV	0.048	0.05	0.345	0.371	0: 0.354
Word Repetition (Tone accuracy) <i>aptitude</i> by <i>test-session</i> by HV+ HVB > LV	0.091	0.082	0.379	0.654	0: 0.758
Training <i>aptitude</i> by HV > LV	-0.037	0.119	0.129	0.572	0 : 0.253
Training <i>aptitude</i> by HVB > LV	0.026	0.101	0.129	0.732	0 : 0.354

The results for *Categorisation of Synthesized Tonal Continua* as the *aptitude* measure are summarised in Table 14. It can be seen that similar to previous results, there is also more evidence for the null than H1 in each case (i.e.  $BF < 1$ ). There was substantial evidence for the null over H1 for Three Interval Oddity task and Word Repetition, tone accuracy. Similarly, in other cases, the robustness regions indicate that even if the scale factor  $x$  was twice as large, i.e. corresponding to the *maximum* value expected, the  $B$  would be ambiguous. However, due to the fact that this task did not positively predict performance on any measure it is not possible to draw strong conclusions from these null results.

Table 14 Bayesian analysis with CSTC as the ID measure, with red cells representing evidence for the Null and yellow cells representing ambiguous results.

Contrast	Mean difference	Stand. Error	H1 estimate $x$	Bayes Factor ( $B$ )	Robustness Region
Picture ID, (Tone accuracy) <i>aptitude</i> by HV+ HVB > LV	-0.304	0.218	0.094	0.623	0: 0.253
Picture Naming, (Tone accuracy) <i>aptitude</i> by HV+ HVB > LV	-0.196	0.146	0.005	0.966	0: 0.152
Three Interval Oddity (Tone accuracy) <i>aptitude</i> by <i>test-session</i> by HV+ HVB > LV	0.046	0.085	0.451	0.298	0.450: $\infty$
Word Repetition (Tone accuracy) <i>aptitude</i> by <i>test-session</i> by HV+ HVB > LV	-0.149	0.139	0.532	0.132	0.202: $\infty$
Training <i>aptitude</i> by HV > LV	-0.273	0.166	0.132	0.498	0 : 0.202
Training <i>aptitude</i> by HVB > LV	-0.247	0.186	0.132	0.482	0 : 0.202

### 3.4 Discussion

This chapter further examines the possible interaction between variability in training materials and individual aptitude using phonetic training of lexical tone. A new training group with multi-speaker input, but controlling for trial-by-trial speaker consistency, was added into the design (high variability blocked condition). This new training condition was compared to the previous HV and LV conditions. The results were highly similar to those reported in the previous chapter: there were no difference between conditions *except* in training and with trained voices in the Picture Identification test. Specifically, in training there was a benefit of LV over HVB and of HVB over HV; with familiar speakers in Picture Identification there was evidence for a benefit of LV over HVB only.

### 3.4.1 Results of perception tasks

The benefit of LV over HVB in training and with familiar voices in the Picture Identification, is similar to the benefit of LV over HV reported in the equivalent analyses in the previous chapter, and reflects adaption to the single speaker used in training in the LV condition. Note that, in training, the benefit does not emerge until the second session, which is predicted since there is one speaker per session so input is identical up to this point. The significant difference between the HV and the HVB group in training, where the HVB clearly outperformed the HV group starting from the first session, suggests that as predicted, and similarly to the finding by Perrichone et al (2011), trial-by-trial consistency in speakers does indeed boost performance in this training paradigm. This is not seen in the Picture Identification, since the intermixing of voices in this test is matched across conditions.

Otherwise, the pattern of results reported in this chapter reflects what has been found in Study 1 and there is no difference between conditions. Critically, this means that despite using the “easier” blocked version of the high variability condition, there was no benefit of encountering multiple speakers during training for generalisation (no evidence that HVB outperforms LV). Importantly, in addition to finding a pattern of null results (i.e.  $p > .05$ ) in the frequentist analyses, additional Bayes Factor analyses collapsing across the two variability conditions also found substantial evidence for the null ( $BF < .33$ ) in all but one of the test measures (Word Repetition, Pinyin, where  $BF = .421$ ). Thus, there is good evidence that, at least for these training and test materials, exposure to stimuli from multiple speakers does *not* lead to greater generalisation in either perception or production. This finding is consistent with the lack of a main effect of *variability condition* in the transfer tasks in either Sadakata & McQueen (2014) or Perrachione et al. (2011). However it is at odds with other phonetic training studies focusing on segmental contrasts (Clopper & Pisoni, 2004; Logan et al. 1991, Lively et al., 1993; Sadakata & McQueen 2013) and with the literature demonstrating a HV benefit in

vocabulary learning (Barcroft & Sommers, 2005, 2014; Sommers & Barcroft, 2007, 2011). This suggests that this overall variability benefit may be restricted to segmental rather than tonal phonetic learning, at least for speakers of a non-tonal L1.

It is difficult to reconcile the lack of benefit for vocabulary learning in the picture naming task, given that vocabulary training experiments by Barcroft, Sommers and colleagues (2005, 2007, 2011, 2014) (as well as Sinkeviciute et al. 2019 for adult learners) have reported benefits of multi-speaker over single-speaker exposure on later picture to word recall. However, one possibility is that this is due to differences in the training set up i.e. 2AFC training focusing on training tonal contrasts, compared with more passive, untargeted exposure in the vocabulary studies. Nonetheless it remains unclear why *tone learning* should be different from other types of phonetic learning in terms of benefiting from speaker-variability. Theoretically speaking, in a framework where all cues compete, variation in idiosyncratic speaker-specific cues would be expected to provide key evidence as to which cues are irrelevant to the phonetic contrast in question (Apfelbaum & McMurray, 2011; Ramscar & Baayen, 2013; Ramscar, Yarlett, Dye, Denny & Thorpe, 2010). This raises the question of how participants in the LV condition are able to generalise at all – i.e. how can they identify the phonetically relevant cues compared with the idiosyncratic cues associated with the single speaker to which they were exposed? One possibility is that other variation in the materials aided generalisation, in particular in the real word stimuli, each tone-contrast is encountered with multiple consonants and vowels. Chen, Qian, Zhou & Guo (2010) carried out time-frequency analysis on 40 different Mandarin syllabic words. The results showed the length of Mandarin tones differs depending on phonetic contexts. For example, tone 1 in /fa/ is generally longer than it is in /yi/. This makes it possible that tonal contrasts with real Mandarin words involved more subtle variation than when tones are superimposed onto English non-words, and this change that may increase both the variability and the difficulty of training. If item variability also aids generalisation to new

speakers, this might explain why there was equivalent generalisation across conditions instead of seeing greater generalisation in the HV conditions (i.e. even the LV condition is actually a HV condition, because of the item variability). On the other hand, Sadakata and McQueen (2014) also saw generalisation even for their LV condition, and in their study this condition lacked variation in terms of both speakers and phonetic contexts. This suggests that the relevant cues for the tone contrasts may be sufficiently acoustically salient for learners to identify them, even when exposure occurs in limited contexts.

Another possibility – and the one suggested by the findings of Sadakata and McQueen (2014) and Perrachione et al. (2011) – is that benefits of high variability for generalisation are masked by individual differences. In their studies, only high aptitude participants showed a high variability benefit, while low aptitude participants did not. It is possible that for lower aptitude participants, the benefits of exposure to varying, idiosyncratic cues are offset by the greater difficulty that these participants have in attuning to the different speakers during training, as discussed above (Section 2.4.5). This explanation is supported by evidence from a study by Goldinger, Pisoni and Logan (1991) who explored the effect of increasing the processing cost of multi-speaker input in the context of word recall (in the L1). Specifically, they exposed participants to single versus multi-speaker word lists, manipulating presentations rates. They found that single-speaker lists produced better word recall than multiple-speaker lists at short inter-word intervals (less than 2000 ms) whereas this effect was reversed for longer inter-word intervals. This suggests that increasing encoding difficulty can remove the benefits of multi-speaker exposure. Relatedly, Sinkeviciute et al. (2019) found that young learners have greater difficulty processing multi-speaker training materials in L2 vocabulary learning, and subsequently fail to show a speaker-variability benefit at test. One interpretation of these findings is that age-related capacity limitations may constrain the ability to benefit from

speaker variability, supporting the notion that differences in capacity limitations can affect an individual's ability to benefit from multi-speaker training.

#### 3.4.2 Results of individual aptitude measures

Returning to the current results, there was no interaction between variability-training and learner aptitude (except for the one which was in the opposite direction found with *Categorisation of Synthesized Tonal Continua* aptitude in the Word Repetition task). However, it is important to acknowledge the results of the Bayes Factor analyses, which did not find substantial evidence in support of the null over H1 (or H1 over H0) for any of the test tasks with *Pitch Contour Perception Test* as the aptitude measure. For analyses using *Categorisation of Synthesized Tonal Continua* as the aptitude measures, there was apparently evidence for the null in the Three Interval Oddity and Word Repetition task. However, the fact that *Categorisation of Synthesized Tonal Continua* does not correlate with measures of baseline performance in these tasks suggests that it is a poor measure of aptitude, so that evidence for the null here is not hugely informative.

The ambiguous results with *Pitch Contour Perception Test* mean that it is not possible to draw conclusions about this hypothesis from the current data. In theory, it is possible to continue collecting data until there was substantial evidence for either H0 or H1. To explore the feasibility of this, supplementary analyses were conducted to estimate the sample size that might be needed to see substantial evidence for the null (based on the assumption that the error term would reduce in proportion to  $\sqrt{SE}$ ). Taking the Picture Identification test (the test most similar to previous studies), the results suggests that it would require  $N > 300$  – i.e. over five times our current sample size. This suggests that this experimental paradigm is not sufficiently sensitive to address this hypothesis.

Given the ambiguity of the current findings with regard to the interaction, it is not appropriate to extensively interpret why the current design did not find the interaction while

the previous studies did. However, there are a variety of differences across the studies which could underpin the different findings, if it holds true. For example, the test of individual differences used is harder than that used by Perrachione et al. (2011) since it uses all six Mandarin vowels (whereas the original study used five, without /y/) and all of the Mandarin tones (where Perrachione et al. (2011) used three, without Tone 3). These changes also mean that that it is difficult to contrast the range of participant scores in the two studies and it may be that the spread of ability of the current participant is different from theirs (the fact that many of the participants had slopes which could not be classified using the *Categorisation of Synthesized Tonal Continua* measure supports this possibility). In addition, the current training task is potentially harder than both of the previous studies, i.e. involving all four tones in the context of natural Mandarin stimuli in the context of a word learning tasks.

### 3.4.3 *Limitations and Future directions*

The results of BF analyses suggest that the current experiment was not sufficiently sensitive to detect an interaction between *variability-conditions* and individual aptitude. One improvement which could be made would therefore be to repeat with a large enough sample to get evidence either for the H1 or the null. Unfortunately, the calculation above suggests that this may be relatively infeasible. Another possibility would be to implement a direct, high powered replication of these previous studies. However again this is likely to require much larger numbers to be well powered, indeed there is an increasing recognition that psychology experiments have been routinely underpowered (Maxwell, Lau & Howard, 2015; and see Vasisht, Mertzen, Jäger, & Gelman, (2018) for a recent demonstration in the area of reading) and that can lead to increases in both type 1 and type 2 error. If a high powered replication were possible, this would also bring verification of the reported results as well as examination on whether the fact that Perrachione et al. (2011) found their interaction with *untrained* voices, whereas Sadakata & McQueen (2014) found it only for *trained* voices, is a true difference (due

to the different paradigms) or due to power. Critically, successful replication could extend the paradigms in such a way as to explore the factors discussed above. For example, would increasing the number of tones to use all four Mandarin tones and/or using natural Mandarin stimuli affect the interaction between variability in the input and learner aptitude?

Although direct replication will play a useful role in establishing these effects, an alternative direction is to develop a more nuanced approach to measuring the factors leading to different levels of aptitude both in tone learning. In addition to not seeing the predicted interaction with *variability-condition*, the current analyses from both Study 1 and Study 2 didn't see a relationship between aptitude and *learning* in any of the tasks (i.e. no interactions with *test-session*). For *Categorisation of Synthesized Tonal Continua* this is likely due to the problems with this measure. However for the *Pitch Contour Perception Test* which *did* predict baseline performance on the task and yet still did *not* show the difference in improvement due to different training stimuli variability. Another limitation of the *Pitch Contour Perception Test* measures is that the task is quite similar in nature to the training and performance measures, decreasing its explanatory value as a measure of individual difference. If a relationship between high variability training and individual aptitude does indeed exist, it may be better captured by individual difference measures of more general cognitive functions. In study 3, I take the step of exploring the predictive value of a range of measures including measures of working memory, attention and musical ability.

## 4. Study 3

### 4.1 Introduction

In Study 1 and Study 2, naïve participants were trained on all four Mandarin tones, using real language stimuli embedded in a word learning task. Improvements were found in both production and perception of tones which transferred to novel voices and items. It was found that during training, performance was the greatest for participants who were trained with a single voice versus four voices (whether intermixed or blocked) but different types of variability in training led to equal amounts of generalisation. Critically, there was no evidence that different levels of aptitude lead to better or worse learning from different types of training input. This contradicts towards previous literature which did find individual aptitude by variability interaction (Perrachione et al., 2011; Sadakata & McQueen, 2014). However, a limitation of Study 1 and 2 is that the individual aptitude measure *Categorisation of synthesized Tonal Continua* turned out not to be a positive predictor of performance in any of the tests. In fact this test did not appear effective with current participants, although a slightly different method was used for calculating participants' slopes than Sadakata and McQueen (2014). Note their measure was used then the majority (>70%) of the participants would have failed to meet their threshold for inclusion in the analysis. The alternative approach – *Pitch Contour Perception Test* – did predict participants' baseline performance in both training and performance measures. Importantly, however, there was no evidence that this measure predicted their ability to learn from the phonetic training materials i.e. it *didn't* correlate with improvement from pre-test to the post-test after training, which would be reflected by an interaction between individual aptitude and test-session.

An additional limitation of using *Pitch Contour Perception Test* as an individual aptitude measure is that it is a direct measure of the ability to identify the Mandarin lexical

tones used in training, and thus it isn't explanatory in terms of which aspects of cognitive ability underpin this ability. However recall in Chapter 1 (Section 1.4.2), tone training studies have also included other measures of cognitive individual differences, specifically measures of working memory (Chandrasekaran, 2010) and musical ability (Li & DeKeyser, 2017). The current study builds on these studies, looking at the relationship between HVPT for Mandarin tones and measures of both musical ability and working memory, as well as measures of attention. There is a substantial literature linking each of these abilities to language learning in general, as well as some linking to learning of L2 lexical tone in particular. Section 4.1.1, (Working Memory) 4.1.2 (Attention) and 4.1.3 (Musical Ability) provide an in-depth introduction to each of these three types of *ID measures*. Each sub-section briefly covers the development of research in this area as well as the evidence linking these cognitive measures to general language ability and lexical tones in particular. Finally, Section 4.2 describes the design choices for Study 3, including the choice of participants (the recruitment of participants currently learning Mandarin as well as naïve learners), the details of the specific tests used to measure individual differences, and changes to the training paradigm compared with previous studies.

#### 4.1.1 *Working memory*

The study of working memory (WM) dates back to around 1960s when researchers proposed a separation between long-term and short-term memory. Studies by Peterson and Peterson (1959) and Brown (1958) demonstrated that there was rapid decay of participants' memory of newly learnt verbal stimuli without rehearsal. This indicated that, in addition to long-term memory (LTM), there is another type of memory system which holds information in a more temporary manner. Atkinson and Shiffrin (1968) suggested that as information was perceived it was stored in a short-term storage system before reaching LTM. The term "working memory" was also introduced to describe this memory system not only as a gateway

to long-term memory, but also as a workspace for complicated cognitive functions such as comprehension. In this original formation of the theory, there was a serial relationship between WM and LTM. Evidence was found in a lesion study by Baddeley and Warrington (1970). They studied a patient with damage to the medial temporal lobes. The patient had severe difficulty in learning new knowledge but performed at average level in WM tasks such as digit span. However, in the same year, another neuropsychological study provided contradictory evidence. Shallice and Warrington (1970) tested a patient K.F. with conduction aphasia. This participant demonstrated clear reduced WM capacity, however, his LTM and other cognitive abilities remained relatively intact. This challenged the model of a simple serial relationship between WM and LTM because if WM is truly the working space for memory formation, then one's LTM can't remain unaffected when WM is damaged. Further data came from Baddeley and Hitch (1974). In their study, typically developed adults were asked to hold a sequence of zero to eight numbers in mind, while performing a variety of tasks thought to require WM involvement from sentence reasoning, language comprehension to learning. Although increased load did impair their performance in these tasks, no dramatic reduction was found. In response to this data, Bradley & Hitch (1974) put forward a multi-component model of WM which is still widely-accepted to the present day. In this model, WM is divided into three separate components which together serve as a working space responsible for many cognitive functions and all contributing to LTM. First is the phonological loop which is responsible for immediate retention of language information such as digits and does not contribute directly to LTM. This could explain the specific deficits found in Shallice and Warrington (1970) - i.e. the patient K.F. suffered from damage to the phonological loop, but not other sub-systems. Second is a parallel system for visual information known as the

visuospatial sketchpad. Finally, the model proposes that there is a central executive which is responsible for attentional controls for these two components.<sup>8</sup>

#### 4.1.1.1 *Phonological loop*

The phonological loop is perhaps the mostly well-researched component of WM. It is believed to consist of two parts: a temporary storage of auditory information which decays quickly and a subvocal rehearsal system that maintain the information within the store as well as retrieving the stored verbal information. There are several types of evidence for a temporary maintenance system that is *phonological* in nature (Baddeley, 2003, 2012). The first is the *similarity* effect. It has been found that when recalling a sequence of items, those sharing more phonological similarity are harder to recall. For instance, letter sequence B, W, Y, K, R, X are much easier to recall than T, C, V, D, B, G, as the latter shares similar sounding names (Conrad & Hull, 1964). Similar findings are found with meaningful words: sequences such as man, cat, map are harder to recall than pit, day, cow. Semantic similarity does not seem to have the same effect (e.g. huge, big, large are no more difficult than old, wet, thin) (Baddeley, 1966a), but instead seems to be more involved in LTM (Baddeley, 1966b).

The second piece of evidence is the *word length* effect i.e. that the recall accuracy decays with the increase in word length (Baddeley, Thomson & Buchanan, 1975). This is believed to be a direct evidence for the rehearsal system: if rehearsal happens in real time, then longer words naturally take longer time to rehearse, leading to more trace decay and worse performance. This can be measured using a non-word repetition task, in which participants are required to memorise and repeat non-words of increasing length. A similar effect is measured

---

<sup>8</sup> In a later version of model, another parallel sub-system, episodic buffer, was added into the model which serves as limited storage of information chunks (Baddeley, 2000).

by another classic test of WM, the *digit span* task, where participants need to memorise and repeat a series of numbers with the length of the series increasing by one number in each trial.

The third source of information for phonological storage is the *articulatory suppression* effect. Baddeley et al. (1975) reported that if participants were required to repeat a single irrelevant word (e.g. the) out loud, then their performance in maintaining a word sequence was impaired and there was no word length effect. In a later study, they showed that articulatory suppression could also remove the similarity effect if the stimuli was presented visually but not if it was presented aurally. This may suggest that there are two sub-systems to the phonological loop: auditory information goes through the phonological store without subvocalisation whereas visual information needs to be subvocalised when registered in the rehearsal system.

#### 4.1.1.2 *The visuospatial sketchpad*

The visuospatial sketchpad is proposed as the visual counterpart to the phonological loop. Evidence for short term storage of visual information comes from studies where printed letters are displayed successively and participants are required to decide if these are the same letter. Posner & Keele (1967) demonstrated identical letters (e.g. AA) are matched more quickly than letters which have the same name but are visually different (e.g. Aa), however this effect disappears with a longer (1.5s) interval between letters. This suggests that the visual information is preserved in a short-term storage for later access but, like phonological information, visual information decays quickly. Later studies avoided use of letter stimuli, which may still allow a role for phonological involvement. Philips and Baddeley (1971) designed a task in which participants viewed a 5x5 square matrix display in which half of the squares were filled randomly in each trial, and then after a 0.3 to 9 - second delay, either the same display was shown, or another display with just one-cell-difference. Participants were required to judge whether the pattern was the same or different. They found that reaction time

increased and accuracy reduced as interval time increased, providing evidence for a rapidly decaying visual memory system. A follow up study suggested that the accuracy of memory also decreased with more cells to be remember (Phillips, 1974).

There has been some debate as to whether the visual-spatial sketchpad is visual-based or spatial-based. Studies generally explore this using different distraction tasks to see which interferes with later recall. Some seem to suggest that spatial information (which could also be auditory) is what is important, rather than visual information (Baddeley and Lieberman, 1980) while others find that purely visual information may play a role (Logie, 1986). It has also been suggested that there may be another movement-based system used in gesture and dance (Smyth & Pendleton, 1990). Regardless of how the different systems interrelate, it is clear that there is a mental process which can be used for immediate retention of material that is not phonological in nature.

#### *4.1.1.3 The central executive and the episodic buffer*

The central executive function is believed to reflect attention-related abilities and thus closely related to Attention as discussed in Section 4.1.2 below. Baddeley (1996) presents this component of WM which is further separated into several components: the capacity to simultaneously execute two tasks (divided attention), the capacity to switch retrieval strategies (attention switch), the capacity to attend selectively to the target while inhibiting distractions (inhibitory attention/selective attention), and the capacity to maintain and manipulate stored information (sustained attention). On the neuropsychological side, there is evidence for a link between the frontal lobe and the central executive. Individuals who suffered damage to the *frontal* lobes often exhibit severe problems in cognitive functional regulations (for an overview of frontal area functions, see Stuss & Knight, 2013), although some studies (e.g. Ahola, Vilkki, & Servo) have found participants with frontal infractions who did *not* have difficulty in

performing tasks relying on central executive (e.g. digit span), and others have found executive dysfunction in patients with *posterior* brain damage (Andrés, & Van der Linden, 2000). Evidence from fMRI research (Collette & Van der Linden, 2002) indicates that multiple *frontal* and *parietal* regions are activated during WM tasks. They argue that central executive is best explained as the interactions between various cognitive functions and multiple brain areas. Some of the decay of executive function observed (e.g. in patients with Alzheimer's disease) may actually be the results of disconnections between these brain areas. There are specific WM tasks which are thought to tap central executive function. For example, Letter number sequencing used in WAIS-III (Wechsler, 1997) requires participants to memorise a series of intermixed numbers and letters and then repeat numbers and letters separately given a specified order (e.g. numbers from small to big, letters with alphabetic order). In order to memorise two sets of intermixed materials, participants need to switch attention between them and maintain attention on one type of stimuli only when needed.

One of the major questions regarding executive function is whether this is one unitary, flexible processes that supports all higher cognitive processes such as reasoning, language and learning, or whether there are multiple systems each responsible for a different set of mental functions. While in Baddeley's model executive function was treated as a single system, in recent years it has been suggested by some researchers that the central executive is a collection of multiple specialised strategies and functions. According to this view, there is no central control involved as these abilities interact and work together towards a self-organising system (e.g. Logie, 2016).

The original model divided WM into information storage (the phonological loop and visuospatial sketchpad) and the control system (central executive function). Baddeley (2000) proposed that there is an extra system which provides further storage for both auditory and

visual information by binding these into “episodes” for access and interacting with multi-dimensional representations in LTM. For example, Baddeley & Wilson (2002) studied the recall of prose passages from amnesic patients. Those participants showed deficits in LTM but could still perform normally on recalling prose passages. As their LTM was impaired, it was assumed that this recall could be attributed to WM. Cowan (2016) has suggested that the capacity for this episodic buffer should be around 4-5 chunks. In the original model, Baddeley (2000) suggested the binding function of episodic buffer (i.e. binding multiple items into episodes) may occur through conscious awareness thus heavily rely on central executive. For visual stimuli, Vogel, Woodman and Luck (2001) has presented an array of objects to participants. After a brief delay, a probe stimulus was used for participants to decide whether this probe had been in the array. Their results suggested that participants can recall up to around four items, regardless of whether there was only a single feature (e.g. colour red) or multiple features (e.g. colour and shape, a red triangle) to remember. Baddeley, Allen and Hitch (2011) looked at this question in more detail. They used concurrent tasks demanding attentional resources which were required for central executive. If binding did require extra attentional resources (i.e. it relied on the central executive), then memorising binding objects (a red triangle) would be affected more by those concurrent tasks. However, no difference between binding or non-binding condition was found. They also designed a condition where an extra stimulus was inserted just before the test, which led to impaired performance. This suggested that binding itself may not require extra attention, but maintaining that information still relies on the central executive. For verbal stimuli, Baddeley, Hitch and Allen (2009) used a similar design to examine this question. In general, sentences were easier to recall than single word lists (sentence superiority effect) as when processing sentence, words are bound into chunks based on syntax, making them easier to recall. Thus, if binding is cognitively demanding then sentence processing will be affected more when other task requiring attentional resources are

performed. With various concurrent tasks, participants' overall performance was impaired however the magnitude of sentence superiority effect remain unaffected. These results indicate that that the episodic buffer may work independently alongside the central executive when combining information.

#### *4.1.1.4 Working memory and language learning*

There is a rich literature on the relationship between WM and language learning. Baddley (1986) emphasised the importance of the phonological loop in learning new words in the L1. In his model, the phonological loop (sometimes referred as verbal working memory) not only stored novel phonological patterns but also worked as a link between short-term and long-term memory for the purpose of learning. Longitudinal studies (e.g. Gathercole, Willis, Emslie & Baddeley, 1992; Gathercole, 1995) with children have revealed large variation in both WM capacity and vocabulary knowledge and that these two are closely related. The strongest evidence was found between performance in non-word repetition tasks and native vocabulary knowledge. However the causal relationship may not necessarily work in the direction of better working memory predicting better vocabulary learning: There is evidence that, non-words sharing similar phonotactic patterns to the L1 (i.e. they are more "word-like") are easier to recall (Gathercole, 2006), suggesting that knowledge of language may also in turn affect participants' performance in tasks measuring WM.

Nevertheless, it is an important question whether similar measures would predict vocabulary learning in L2 learning. Early work by Service and Kohonen (1995) showed that non-word repetition using pseudowords with English phonology predicted better English scores in Finnish primary school children aged 7-10 who were studying English at school. Cheung (1996) used 12-year-old Hong Kong students studying English as an L2 and again found that their English non-word repetition span predicted their English learning ability,

although this relationship was only found in those whose vocabulary size was relatively small. However, these studies all use measures of non-word repetition and stimuli with the L2 phonology. Speciale, Ellis and Bywater (2004) recruited English speakers and taught them German new words in one experiment session and also looked at English speakers learning Spanish in a 10 week course. They measured WM with a non-word repetition task using words which were judged to be different from both L1 and L2 vocabulary, and found that this predicted L2 performance in each case. Looking beyond vocabulary learning, Kormos and Sáfár (2008) looked at Hungarian secondary school students learning English and the relationship with performances in non-word repetition and digit span backward tasks carried out in the L1 (i.e. Hungarian non-words and digits). English performance was measured by the high school final exam, including assessment on vocabulary, grammar and language comprehension. Both working memory measures were predictive of English performance, although non-word repetition only predicted the performance for those who started at a relatively high level at the beginning of the year.

Turning to the topic of the current thesis, is there any reason to believe that WM will relate to the learning of lexical tone? For non-linguistic tone, Strait, Kraus, Parbery-Clark and Ashley (2010) found that auditory working memory, as measured by a digit span task, predicted individuals' ability to discriminate non-linguistic tones with different frequencies: Individuals with higher digit span measure were able to distinguish a larger range of tone frequency. George & Coach (2011) also reported a positive correlation between digit span task performance and the efficiency of differentiating artificial tones (800Hz vs 820 Hz). Turning to the lexical tone used in natural languages, Ou, Law and Fung (2015) as well as Ou and Law (2017) looked at factors predicting the perception and production of tone contrasts for Cantonese speakers in their native language. Specifically, they studied tone contrasts T2/T5

and T4/T6 contrasts, which are known to cause difficulty even for native speakers. They used a battery of working memory tasks (WAIS-IV, Wechsler, 2008). They found evidence that perception performance (RT in a discrimination task) was predicted by working memory tasks, although surprisingly this was true for visual working memory tasks but not auditory working memory tasks (although auditory *attention* tasks were predictive in production - see section 4.1.2.1). In terms of the relationship between WM and tone learning in an L2, as discussed in Section 1.4.2, Chandrasekaran et al. (2010) conducted a training study in which participants learned lexical tone in the context of a novel word learning task and they examined whether this learning in this study was related to auditory WM as measured by a digit span task (Digit Span Backwards – described in section 4.2.1 below) and Letter Number Sequencing tasks. They did *not* find any evidence that these abilities were related to tone learning ability.

To conclude, there is substantial evidence that WM is related to language learning, including L2 learning. There is also some evidence that WM is linked to the processing of both non-linguistic and linguistic tones. However, the one study which investigated whether WM was connected with the ability to *learn* lexical tones via phonetic training did *not* find evidence for this. However, as discussed in Study 2 (Section 3.4.3), it is important not to over interpret null results as the relationship may be due to Type 2 Error. Given the substantial literature relating WM with phonological language learning, the current study aims to further probe the relationship between lexical tone learning and WM using a variety of measures: *Digit Span Forward* and *Digit Span Backward* tasks, which measure both auditory WM capacity and the function of the phonological loop, the *Letter Number Sequencing* task which captures the function of central executive and *Arithmetic* as a measure of general WM and cognitive function.

#### 4.1.2 Attention

The processing of multiple cognitive functions relies on the attention system as the central control. The allocation of attentional resources is also very important in almost all daily scenarios. The most widely used attention model was put forward by Posner and Petersen (1990). They posit three main networks involved in the attention system: an alerting system responsible for sustained vigilance, an orienting system responsible for directional information and an executive system as the control. There is a range of neuropsychological and behaviour evidence in support of these systems, which will be discussed below.

In general, there are two types of alertness generated by the human system: tonic alertness which represents the intrinsic arousal that fluctuates over time, and phasic alertness which represents the immediate, rapid change in attention caused by a brief event. The tonic alertness is believed to contribute to sustained attention and provides the cognitive resource for higher cognitive functions such as WM, while phasic alertness may be mainly involved in more short-term cognition such as selective attention. Early research focused on the alerting system in terms of arousal (tonic alertness). Using animal brains, studies have shown that the brain stem reticular system is highly involved in maintaining alertness level (e.g. in cats: Graybiel, 1977; in monkeys: Keizer & Kuypers, 1989). Later, arousal was further defined as the ability to create and maintain optimal vigilance and performance when necessary. This ability is important as keeping an appropriate alertness level is beneficial for processing signals with different priorities. Posner (1978) looked at the difference between tonic alertness and phasic alertness in humans. Using letter and word matching tasks, (similar to the one used by Posner and Keele (1967) as described in Section 4.1.1.2) he found that for a familiar letter, the passive activation (tonic alertness) happened a constant rate. This passive activations was believed to reflect the processing of the basic information of the item such as its name, physical form and

its semantic classification (e.g. vowel or not a vowel). Thus, if the matching criteria was based on these aspects, alertness level did not affect participants' RTs. The phasic alertness level, as induced by showing a warning signal before the target appeared, however, did affect the response speed when more complicated information was processed (e.g. a new word was used) with a higher alert state resulting in quicker response but higher error rate. These results suggested that the accumulation of information about the target was independent of phasic alertness level. Higher alerting state affected the speed at which the attention system could respond to a stimulus, but did *not* allow participants to make more accurate judgements. Thus, participants would have less time to accumulate information resulting in worse performance. Neuropsychological study also provided further evidence regarding the difference of alertness types. Cerebral blood flow has revealed heavy involvement in right brain hemisphere when participants undertake vigilance tasks (e.g. respond quickly to emerging visual signals) for tonic alertness (Cohen et al., 1988). In addition, there is evidence that lesions of the right hemisphere can cause alerting difficulties. For example, Yokoyama, Jennings, Ackles, Hood and Boller (1987) have reported that, unlike healthy participants, individuals with right cerebral lesions did not show any change of heart-rate when faced with warning signals. Using PET and fMRI, Sturm & Willmes (2001) reported a right-hemispheric frontal, parietal thalamic and brain-stem network for both tonic and phasic alertness. For phasic alertness only, there was extra activation of left- hemispheric frontal and parietal areas. They interpreted these activation patterns as the effect of selective attention. This was supported by a later fMRI study (Fan, McCandliss, Fossella, Flombaum & Posner, 2005) which found evidence that the warning signal effect relied on left hemisphere activation. Finally, there is evidence that one of the neuromodulators- norepinephrine (NE) - may be the physiological foundation of the alerting system. This is released by the locus coeruleus in the brainstem which is also activated when participants were presented with warning signals. Using drugs which block or increase NE

release has been found to diminish or amplify the effect of warning signals (Aston-Jones & Cohen, 2005).

The orienting system is mainly responsible for processing the sensory information such as visual location and pattern recognition. Studies have shown that this orienting system can work independently of the alerting system. While the alerting system mainly focuses on the “when” aspect of perception, the orienting system mainly focuses on the “where” aspect. Beane and Marrocco (2004) designed a task to separate information about “where” and “when” with four different cue-conditions. There was a no-cue condition as a baseline, as well as conditions where participants were either only informed *where* the target would appear (orienting system), or *when* the target (alerting system) would appear, and a condition where both cues were presented. In this way, by comparing across the different conditions, they could isolate the effects of each cue individually. Results suggested that the two functions were affected by different chemical mechanism. The orienting system was mainly controlled by the neuromodulator acetylcholine while the alerting system was mainly controlled by the neuromodulator NE as suggested above. There is also evidence suggesting that the performances on these two systems are not correlated (Fan, McCandliss, Sommer, Raz & Posner, 2002). These findings suggest that although in daily life the information regarding time and location are processed at the same time, meaning these two systems must work closely with each other, they are nevertheless independent and separated systems. In terms of neural underpinnings, some studies have indicated that the frontal eye fields (e.g. Thompson, Biscoe & Sato, 2005) and parietal areas (Lindner, Iyer, Kagan & Andersen, 2010) are involved in the orienting process. These areas are also believed to be part of the dorsal pathway which is responsible for processing spatial location and provide guidance for actions. Meanwhile, the temporoparietal junction and ventral frontal cortex are involved when there was a need to shift

attention. The synchronization between the dorsal and ventral attention may be the key for the orienting system to function (Womelsdorf, et al., 2007).

The executive control system has been suggested to arise from the limited capacity of attention system. The ability to attend to different objects when necessary is also called focal attention. It has also been argued that the use of executive control is the entry to the conscious state which provides top-down regulation over the system. There is evidence for the involvement of midline cortex and the anterior cingulate cortex (Dehaene & Changeux, 2011). Further study has also suggested another executive control network located in the fronto-parietal brain areas. While the original cingulo-opercular system acts to provide stable background maintenance for the overall performance, the fronto-parietal system works on a trial basis to initiate, switch and adjust attention in real time. Although this is similar to the dorsal pathway as described in the orienting system, it is believed to be independent from it (Dosenbach, Fair, Cohen, Schlaggar and Petersen, 2008). An alternative theory posts an overall cognitive control model (Carter & Krug, 2012) in which the lateral prefrontal cortex provides top-down control signals and this is modulated by performance-monitoring signals from middle brain. Recall from Section 4.1.1.3 that WM is also believed to be modulated by a central executive control component and posterior (Andrés, & Van der Linden, 2000), frontal and parietal brain areas may be involved (Collette & Van der Linden, 2002), which overlaps with the executive system described in attention models. The nature of the central executive function is still under debate. While some suggests that both WM capacity and executive function is based on a common attention component (McCabe, Roediger, McDaniel, Balota & Hambrick, 2010), other suggest that executive control emerges from the interaction between brain areas involved in WM and attention (Gruber & Goschke, 2004). Either way, both WM and attention

are likely to be controlled by some executive function which may result from the cooperation of various brain systems.

More recent study has also explored the development of this attention system. An fMRI study by Posner, Rothbart, Sheese & Voelker (2012) has reported that although in adults these systems are believed to work independently, for *infants*, the orienting system also provides executive control, suggesting that these systems may not be entirely separated in the early stage of life. There is also evidence that the functioning of these systems may exhibit great individual differences. For example, the DRD4 gene has been reported to affect the efficiency of the executive system. Children with a specific mutation type (DRD4 7-repeat allele) may be more susceptible to the surrounding environment, making it hard for them to hold attention and possibly leading to ADHD (Bakermans-Kranenburg & Van Ijzendoorn, 2011). Adults with such a mutation may exhibit weaker attentional control, making them, for example, more vulnerable to alcohol addiction in certain environments (Larsen et al., 2010). Attentional training such as meditation has also shown to be effective in improving executive control and changing corresponding brain areas, suggesting that attention is a dynamic system that can be trained (Rueda, Rothbart, McCandliss, Saccomanno & Posner, 2005; Tang, et al., 2009).

There is also a literature on the relationship between auditory and visual attention, and how separable these are. Note that the integration of auditory and visual attention is crucial for various real-life scenarios ranging from language learning (Norrix, Plante & Vance, 2006) and lip-reading (Pekkola et al., 2006) to movement recognition (Bidet-Caulet, Voisin, Bertrand & Fonlupt, 2005) and locating targets in complex environment (Best, Ozmeral & Shinn-Cunningham, 2007). Both auditory and visual attention mechanisms rely on both the top-down processes and the bottom-up processes. Top-down attention is important for allocating cognitive resources to the most appropriate sensory input to avoid the effect of distraction. This

attention leads to increased sensory sensitivity, shorter response times and more accurate information processing (Sussman, Winkler & Schröger, 2003). On the other hand, bottom-up processing is more automatic. For auditory attention, it is auditory saliency based (Kayser, Petkov, Lippert & Logothetis, 2005) while for visual attention, it is image salience based (Itti & Koch, 2000). Also, as described above, the orienting system is one of the fundamental attention systems. Both auditory and visual perception need to attend spatial information, and spatial attention itself is thought to be supramodal. For example, the auditory ERP for auditory stimuli was enhanced with extra visual spatial cues (Hötting, Rösler & Röder, 2003).

There is evidence that both auditory and visual attention draw cognitive resources from the same pool. Not only does attending to auditory stimuli enhances the activation in auditory cortex, activation is *decreased* if there is an obvious visual distraction, and vice versa for visual cortex activation (Slevc, & Miyake, 2006). This leads to a limited-capacity model in which there is only limited attentional resources, thus attending to one type of stimuli necessarily reduces attention to other (Lee & Faber, 2007). The ability to switch attention between different types of stimuli seems to rely heavily on the dorsolateral prefrontal cortex since participants' ability to switch between irrelevant visual and auditory stimuli is impaired if transcranial direct current stimulation was applied to this area (Nikolin, Martin, Loo & Lauf, 2018). Another similar process is selective attention, where participants needed to attend to only one type of stimuli but ignore the distraction from the other. There is evidence that this process involves the use cognitive control to increase the attention to relevant stimuli or decrease the attention to irrelevant stimuli. More important, if more distracting stimuli are presented, this compensation effect also gets stronger (Weissman, Warner, & Woldorff, 2011). Later studies have suggested that the attention system does not allocate resources to all input but only to those that *might* be relevant. For example, if the auditory sound “right” was the distractor and

the task was to identify the stimuli on screen, in this case displaying the left arrow would be less affected comparing with displaying the word “left”, as individuals assign no or limited attentional resources to a distractor which does not share similar format with the target (Grant & Weissman, 2017). However, it should also be noted that, although it seems that auditory and visual attention may “compete” under limited cognitive resources, it is unclear whether for more basic tasks such as a pitch discrimination task or a visual contrast discrimination task, there will be such a conflict. Alais, Morrone and Burr (2006) required participants to process two such tasks simultaneously and found no deterioration in their performance in either of the visual or auditory tasks compared with when they were performed independently. Thus, in some contexts, at least for typically-developed adult, cognitive resources may be sufficient for processing and visual and auditory attention do not necessarily compete with each other (although again it is important not to over interpret null results in the literature).

To summarise, similar to WM, attention also functions under limited cognitive resources. With the integration of both visual and auditory attention, higher cognitive functions such as language learning becomes possible. The evidence for a relationship between attention and language learning is reviewed in the next section.

#### *4.1.2.1 Attention and language learning*

As introduced above, research in attention has mainly emphasised the ability to focus on and encode relevant information, despite the existence of simultaneous distracting signals. Developmental research has provided evidence that the ability to selectively tune to a target language begins early in infancy. Lalonde & Werker (1995) studied infants between 8-10 months. Previous work had shown that by this age, infants have already developed biases based on their native language, such that they can discriminate native contrasts but not non-native contrasts. They were interested in how this behaviour interacted with developing aspects of

cognition. The results suggested that infants' ability to discriminate non-native tones correlated with their performances in visual categorisation and target search tasks, and this correlation could not be explained by a simple age effect. Thus, they concluded that such cognitive competencies may influence speech perception development by the end of 1<sup>st</sup> year of life. This inhibitory control process is also seen as a domain-general ability and it occurs regardless of the type of native languages or the stage of language development (Conboy, Sommerville & Kuhl, 2008). The study of developmental disorders such as autism spectrum disorder (Tye et al., 2014) and developmental dyslexia (Vidyasagar & Pammer, 2010) has revealed that patients suffering from these disorders exhibit damaged attentional ability. Tallal & Piercy (1973) found that dyslexic children were able to discriminate two non-verbal tones with different frequencies but failed to do so if these two tones were presented with an interval less than 400ms. Hari and Renvall (2001) reviewed other research looking at both verbal and visual stimuli processing and concluded that the cause of this deficit was a slowed attention shift such that dyslexic children took significantly longer time to shift their attention to the next target. More recent research has confirmed this using eye tracking: Facoetti et al. (2010) found that typically-developing children have quicker attention orientation to both visual and auditory stimuli compared with dyslexic children, and dyslexic children can't disengage from the stimuli efficiently. These studies indicate the importance of attention switching abilities in discriminating between auditory stimuli. There are similar findings with dyslexic adults, for example Lallier et al. (2010) found that these participants required a much longer inter-stimulus interval in order to process two successive auditory stimuli compared with controls, suggesting longer time is needed to finish the attention shift. This suggested that individuals have difficulties in reassigning and switching auditory attention between items making it difficult to process successive auditory stimuli.

For non-impaired adults, there is relatively little research on the role of attention within their native language (e.g. Native English speakers' performance on English). Andrews (2012) suggests that this may be due to an implicit "uniformity assumption" in linguistic research which believes the ability of processing one's native language should be relatively constant across all individuals. However, this idea has been challenged in recent years particularly within the literature on reading. Veldre and Andrews (2014) looked at skilled reading employing a gaze-contingent and moving-window paradigm. In this task, participant reads some text and letters in the sentences are replaced with "X" around the target word to examine participants' use of parafoveal information. Results suggested that although all participants were skilled readers (demonstrated by 95% comprehension rate of all trials), there was significant individual differences between their reading speed, fixation duration and saccade length. In addition, they reported participants with higher reading abilities also have higher sensitivity towards upcoming information. The authors explained these differences both in terms of better lexical retrieval and also importantly, attention switching abilities, although they did not measure this directly.

More direct evidence for a role for attention is found for adults who are learning an L2. For instance, Hazan and Kim (2010) trained British English speakers to discriminate Korean alveolar lenis /t/ and aspirated /th/ stop. Although participants' learning rates did not correlate with some cognitive measures such as WM, their measure of selective attention *did* correlate with their ability to learn the phonetic contrast, as measured by their improvement during training. Other studies have looked at L2 learners with different levels of proficiency. Díaz, Baus, Escera, Costa and Sebastián-Gallés (2008) studied Spanish-Catalan bilinguals. Their Catalan proficiency was measured by assessing their ability to perceptually differentiate the Catalan /e/-/ɛ/ vowel contrasts, dividing them into good perceivers and poor perceivers. Their

learning was assessed via their perception of both native (/o/-/e/) and non-native (/o/-/ö/) phonetic contrasts. Using an ERP paradigm, they found a difference in mismatch negativity (MMN) was correlated with participants' perception accuracy of these contrasts, such that poor perceivers showed decreased MMN compared with good perceivers. The MMN reported in their design was attributed to the superior temporal and a frontal generator, the latter of which is believed to capture an involuntary attention switch toward detecting change in the auditory input.

Turning to L2 tone learning specifically. Lin and Francis (2014) provide evidence that native speakers of a non-tone language (English) and speakers of a tone language (Mandarin) differ in how they assign their attention to consonants and tones. Participants undertook a 2AFC task in which they heard a stimuli in their native language and had to categorise it in terms of vowel, consonant or "tone". Tones 2 and 4 were used for Mandarin and intonation patterns corresponding to the symbols "?" and "!" were used for English. Stimuli were constructed in group sets so that in the baseline set, only the target dimension in the trial varied (e.g. tones varied while the vowel and consent remained stable, /pi/ T2 and /pi/ T4). In the orthogonal set, the values of the non-target dimension varied independently of the target dimension, while the third dimension remained constant (e.g. both tones and vowels varied independently but the consonant remained unchanged, /pi/ t2, /pai/ t2, /pi/ t4, /pai/ t4). They found that Mandarin speakers showed slower response in the orthogonal condition, suggesting they processed consonants and tone in a combined manner, while English speakers was not slowed down suggesting they treated them perceptually separable. Further evidence was provided that even when Mandarin speakers were taking the test in English version (i.e. they heard English stimuli and had to select "?" and "!" rather than tone 2 and tone 4), they still demonstrated slower response in orthogonal condition. This suggests that even in this condition their experience of

Mandarin leads them to engage attentional resources since they naturally attend to both type of information, even though the task only requires them to discriminate one type of information. Thus, the key to learning Mandarin may lie within the reallocation of attentional resources. Zou, Chen and Caspers (2017) found similar evidence comparing native Dutch speakers with no Mandarin experience against experienced learners and native Mandarin speakers. Here they used Mandarin stimuli with all groups. Specifically, participants heard Mandarin CVCV non-words accompanied by Mandarin tone 2 or tone 4. They were tested on their discrimination of tones using an ABX task (i.e. deciding which of the first two stimuli matches to the third one) and compared the following conditions: *Forced-segment*: participants must match X based on segment (A & B differ in segments but with the same tone), *Forced-tone*: participants must match X based on tone (A & B differ in tones but use the same segments), *Segment-and-tone*: participants can match X based on both tone and segments, *Segment-or-tone*: X shares matched segments with A and matched tone with B, so participant must choose which cue to attend to. As would be expected, in the *forced tone* condition, Mandarin speakers and advanced learners were more accurate than naïve Dutch speakers and beginner learners. Similarly, in the *segment-and-tone* condition, native Mandarin speakers were better and faster than the other three groups. This suggests that the ability to identify a match using tone grows with experience of the language. Critically, however, in the *forced segment* condition, while there were no difference between groups in terms of accuracy (which was high overall), Mandarin speakers and advanced learners responded significantly slower than naïve Dutch speakers and beginner learners, suggesting that they have more difficulty not attending to the tones. Similarly, in the *segment-or-tone* condition, Mandarin speakers and advanced learner chose X on basis of tone significantly more than naïve Dutch speakers and beginner learners and, importantly, advanced learners were significantly slower than naïve Dutch speakers and beginner learners, suggesting again that greater experience with tone may lead to greater competition between recourses.

Overall, it seems that learning Mandarin causes more cognitive resources to tone than for naïve participants, potentially leading to greater accuracy but slower processing.

More direct evidence that measures of attention correlate with the perception of lexical tones comes from the work by Ou, et al. (2015) and Ou and Law (2017) discussed in section 4.1.1.4. Recall that they tested native Cantonese speakers' perception and production of Cantonese T2/T5 and T4/6 contrasts and compared this the results of a battery of cognitive tests. Specifically relevant here is that they measured their attention switching and sustained attention using the Test of Everyday Attention (TEA). Results showed that attentional switch ability predicted tone accuracy in both speech perception and production while sustained attention measure was not predictive of neither perception nor production.

Taking all these findings together, it has been seen that attention plays an important role in phonetic training for non-tonal contrasts, and there is evidence that experience learning a tonal language changes the allocation of attention between tones and other phonetic segments. Also, for native speakers, individual differences in attention are related to variation in the perception of different tonal contrasts. However, to date, there is no direct study looking at whether attention is predictive of the ability to learn lexical tones in a training paradigm. The current study aims to look at this in a phonetic training study using TEA to assess various aspects of participants' attention including both sustained attention and attention switching ability.

### *4.1.3 Musical ability*

The study of musical ability as an aspect of cognition began in the second half of the 20<sup>th</sup> century. Early work focused on developing tests to quantify musical ability. Many of these focused on aural skills such as music tone differentiation (for a review, see Shuter-Dyson,

1999), although some studies also explored musical ability in terms of “musicality”, emphasising the aesthetical recognition of musical ability as measured via self-reported questionnaires (Lundin, 1953). Other work suggested that musical ability comprised a more complex set of abilities: Seashore (as summarised in Seashore, Lewis & Saetveit, 1956) argued that there was no such thing as “musical ability” in general, and that it should instead be divided into a set of discrimination skills over different types of sensory information - pitch, timbre, rhythm, loudness, time and tonal memory, which he argued depended on different biological functions and were unrelated. He also suggested that these abilities were largely pre-determined and stable over time. In contrast, Wing (1970) proposed that musical ability should be treated as a uniform, general ability to perceive and appreciate music and designed a set of tests to capture musical intelligence, for example a test of tonal memory capacity, although these tests have been criticised as incomplete. For example, Lynn, Wilson and Gault (1989) suggested that these tests neglected participants’ processing speed. Gordon (1979, 1982, 1989) developed a series of different tasks aiming to capture musical ability from pre-school stage to adulthood. He developed primary, intermediate and advanced measures based on cultural background, tonal ability, rhythmic ability and musical sensitivity. While Gordon’s tasks mainly used natural sounds of music instruments, some other researchers have used synthesised stimuli which allows more subtle tests of identification and discrimination ability. For example, Vispoel (1993) created stimuli varied in melody length, frequency length and size, position and direction of note change. The results suggested that these synthesised stimuli provide similar aptitude measure results compared with traditional methods, suggesting that the music perception is underpinned by sensitivity to the frequency and temporal nature of sound patterns. A further key area of debate has been the extent to which musical ability depends on “nature” versus “nurture” – i.e. the extent to which music aptitude is biological or dependent on relevant

musical education. Studies has suggested there may be a biological component in musical expertise but this ability can certainly be trained with practice (Hallam, 1998, 2004, 2010).

#### *4.1.3.1 Musical ability and language learning*

There is a large literature looking at the relationship between musical ability and language ability. To begin with, similar to language abilities, the ability to understand and process music is seen as a unique function of humanity (Patel, 2006). It has been pointed out that both language and music have a kind of syntactic structure (Patel, 2003). There is also evidence that there are overlaps between brain areas associated with language and music processing. For instance, harmonic progression which requires participants to process musical syntactic information (identify unexpected musical event in chord sequences) activates Broca's and Wernicke's areas, key areas known to be involved in language processing (Koelsch et al., 2002). Moreover, speech sounds and musical pieces are found to be processed similarly in the auditory system: FMRI experiments (Schön et al., 2010) revealed the bilateral activation of middle and superior temporal gyri and inferior and middle frontal gyri when listening to spoken words, vocalisation (singing without words) and sung words. The authors suggested that there may be a more common cerebral network for both phonological and melodic processing. A review by Patel (2010) postulated a universal hierarchical structure for both language and music and that, although these different types of information may be stored in different brain areas, there may be a common neural network for interpreting the structure of music and speech sounds.

Previous research also compares the development of music in the early stages of life with the development of language. In typical development, the ability to produce and discriminate tunes is developed effortlessly, much like basic language skills (Trainor's, 2005). Another similarity between music and language is that in both cases there seems to be a

relatively stable, fixed order in which abilities emerge. For musical ability, infants begin by being able to distinguish certain frequencies and melodic contours (Trehub, 2001) and then later develop ability to discriminate more complicated pitches and tonalities (Krumhansl & Keil, 1982; Trehub, Bull & Thorpe, 1984). The development of language also exhibits a relatively stable chronological order. For example, across languages, during the first year of life infants develop sensitivity to the phoneme systems of their native language (Werker & Tees, 1984) and begin babbling in production (Locke, 1989). Later, they produce first word around 12-month-old and begin form sentences around 24-month-old (Luinge, Post, Wit & Goorhuis-Brouwer, 2006).

There is also a literature on the relationship between musical ability and speech processing in L1. Kraus, Strait and Parbery-Clark (2012) measured speech-in-noise perception ability, auditory cognitive skills and speech-evoked brain stem activities in trained musicians (adults and children) and their peers. For musicians, the three measures were highly correlated with each other. This well-tuned auditory system for music may in turn benefits their neural and cognitive mechanisms for language. Parbery-Clark, Tierney, Strait and Kraus (2012) reported that individuals with musical experience could better discriminate speech sounds with small differences (/ga/, /da/ and /ba/). There is also evidence suggesting a relationship between musical ability and high-level language functions (syntax) processing in children (Jentschke and Koelsch, 2009).

Turning to the relationship between L2 learning and music experience, an influential study by Slevc and Miyake (2006) explored the relationship between musical ability measured using Wing Measures of Musical Talents (Wing, 1968) and four aspects of L2: receptive phonology, productive phonology, syntax and lexical knowledge. Participants were native Japanese speakers living in the US, with time in the country ranging from 6 months to 25 years.

Although only a weak relationship between musical ability and syntax or lexical knowledge was found, results revealed a clear link between musical ability and L2 phonological ability, both receptive and productive. Particularly relevant to the current thesis is the line of research exploring the relationship between musical ability and learning pitch patterns (tones). Studies have revealed that musicians are particularly good at identifying pitch variations: Micheyl, Delhommeau, Perrot and Oxenham (2006) found that, for pure and complex tones, musicians' pitch discrimination thresholds were six times shorter than non-musicians, indicating they could distinguish much smaller differences. Relating this to language, Marques, Moreno, Luís Castro & Besson (2007) studied whether French native speakers were sensitive to changes to the pitch in final words of sentences in an unfamiliar language (Portuguese) and found that musicians were better and quicker at discriminating pitch deviations than non-musicians. Turning to lexical tone, there is evidence that this sensitivity to general pitch patterns may extend to linguistic pitches in tonal L2s, such as Mandarin. Some early studies reported a potential advantage of music training on Mandarin tone identification and discrimination such that native English speakers who have received musical training can discriminate and identify the four Mandarin tones with better accuracy, however these had small samples (e.g. Gottfried & Riester, 2000, 7 musicians with 35 participants in total; Alexander, Wong & Bradlow, 2005, 9 musicians with 18 participants in total). Gottfried, Staby & Ziemer (2004) showed that musicians (Native American English speakers) outperformed non-musicians in producing the four Mandarin tones. In a follow up study, Gottfried and Ouyang (2005) suggested that, in particular, these musicians pronounce Tone 4 better than non-musicians as measured by F0 similarity with native Mandarin speakers. A later study by Lee and Hung (2008) examined the ability to identify tones in terms of the separate identification of pitch height, pitch contour and pitch variability. They found that the advantage for musicians mainly came from their accurate identification of pitch contour. Another study by Musacchia, Strait & Kraus (2008) suggests

that a possible neural underpinning of this benefit might lie within the brainstem response which reflects activity in the auditory nerves, since these responses were generally stronger for musicians than non-musicians when presented with speech or music stimuli. Wong, Skoe, Russo, Dees and Kraus (2007) further reported that when participants were presented with Mandarin tones, a stronger representation was formed in the brainstem auditory responses for Musicians compared with non-musicians, highlighting the possibility that Mandarin tones are processed in similar way as musical tones. This was found even in those musicians who did not know Mandarin tones. Further evidence that musical ability is related to the processing of tones comes from Delogu, Lampis and Belardinelli (2010) who conducted an experiment with adults and children who had no previous experience of any tonal language. They employed a same-different task assessing participants' ability to detect phonological and tonal variations in pairs of monosyllabic Mandarin Chinese words. They also measured their melodic proficiency using a Perceptive Tonal Memory Test from Wing Measures of Musical Talents (Wing, 1968), in which participants hear two melodies differing in only a single tone, and are required to identify whether two melodies are the same or different. They found that both adults and children performed better on the trials where they had to detect differences in phonology than in trials where they had to detect differences in tones, however their melodic proficiency was a good predictor for tonal, rather than phonological trials.

Most relevant to the current study is the training study by Li and DeKeyser (2017), which was discussed in Chapter 1 (Section 1.4.2). Recall that they trained native English speakers, with no previous experience of any tonal language, on Mandarin tone words, using either productive or perceptive training. They measured musical tonal ability by combining performances on tests of pitch perception ability, tone differentiation ability (both from Wing Measures of Musical Talents, Wing, 1968) and the ability to reproduce a list of tones (2-7 notes

long) (Slevc & Miyake, 2006). This was correlated with both overall tone-word perception accuracy and overall production rating regardless of whether participants were trained using production or perception.

In conclusion, there is a scientific field of music cognition with a large literature exploring the nature of musical ability, and there are many parallels between music cognition and language. There is evidence that music is related to various different aspects of language such as syntax and speech processing. Most relevant to the current thesis, there is also direct evidence that musical ability is related to the processing of tones, including a previous training study where they found that musical tonal ability was predictive of both perception and production performances in Mandarin. The current thesis will further investigate whether musical ability is predictive of the ability to learn lexical tone in HVPT training paradigm.

#### ***4.2 The current study***

The current study aimed to further explore the relationship between individual differences and Mandarin tone learning with high variability materials. In previous literature (Perrachione et al., 2011 & Sadaka & McQueen, 2013), they found that the ability to identify and discriminate Mandarin tones interacted with the type of training stimuli used when learning Mandarin tones, that HV materials only benefitted participants with high aptitude. In Study 1 and 2, the experiments looked for a similar pattern using very similar aptitude tests to the two previous studies, although using these as continuous rather than categorical predictors. Specifically, it was predicted that if high aptitude participants benefit more from high variability materials (as found in the previous studies), there should be a stronger positive correlation between the measures of aptitude and the measures of learning, i.e. the change in performance from pre- to post- test would be larger for HV compared with LV materials.

However, neither of the experiments found this pattern. In fact, overall there was no evidence that either of the aptitude measures was predictive of pre-to-post improvement for any test. The current design focuses on HVPT specifically and aims to further examine the role of individual differences looking at a range of measures of WM, Attention and Musical Ability. The *Pitch Contour Perception Test* used in Study 1 and 2 is also included. In the current study, this is used both as a measure of individual difference (looking just a performance on this test at pre-test, as in Study 1 and 2) and also as an outcome measure (looking at improvement from pre-to post-test) with the consideration that previous results suggested that performance on this task improves after training (Section 3.3.2.1).

As discussed above, it appears that the processing of lexical tones might be different for naïve individuals with no experience of tonal languages compared with those who already acquired some experience of the language (e.g. Zou et al., 2017). This raises the question whether to look at training in naïve participants or those who already have some experience with the language might make a difference. Both Study 1 & 2 used naïve participants in line with previous lexical tone training studies. However, outside of tone training, most HVPT have used current learners of the L2, and in terms of educational relevance, current learners are population for whom this type of training would likely to target. Recall that Hao (2014) reported that some tone contrasts are likely to be difficult even for advanced learners, so it is interesting to see if this training is effective even for those with some experience. There is also evidence that different cognitive factors may be relevant at different stages in learning a language (Díaz et al., 2008; Kormos & Safar, 2008). Therefore, the current study includes and compares both naïve participants and current learners. Another advantage of this approach is that is by looking across the two groups together, it is more likely see a wider spread of

performance in the training and performance tests, which could make it easier to find relationships with individual differences.

The remainder of this section first provides some details about the choice of tasks used to measure WM, Attention and Musical Ability, followed by a description of the training paradigm used and how it differs from that used in Study 1 and 2.

#### 4.2.1 *Measures of individual aptitude*

For WM measures, a weakness of previous studies is that they use a single task to measure it. Many studies simply used digit span task (e.g. Pisoni & Cleary, 2003) or a non-word repetition task (Speciale et al., 2004) as the measure of WM. However, although these tasks often considered classic measures of general WM capacity, they may not reflect all the aspects regarding the phonological WM used in language learning. For example, these tasks capture phonological loop storage but not central executive function. Here, I used the Working Memory Index (WMI) from WAIS-III test (Wechsler, 1997). This is a test set which has been widely used across the world (for an independent assessment on the reliability on WAIS-IV which used a same Working Memory Index as WAIS-III, see Benson, Hulac & Kranzler, 2010). The tasks involved are *Digit Span* (forward and backward), *Letter Number Sequencing* and *Arithmetic*.

I noted above that Digit Span tasks require participants to recall and repeat a series of numbers with the length of the series increasing by one number in each trial. In the current battery, this is the *Digit Span Forward test*. In the current study I also include a common variation of this task which is the *Digit Span Backward test*, where participants need to repeat the memorised numbers in reversed order. This more difficult task may capture more subtle differences in phonological working memory. Including both tests follows the recommendation

of a recent review of the Wechsler Intelligence Scale by Weiss, Saklofske, Coalson and Raiford (2010). They found using only digit span backward task resulted in a decrease in performance, even for high-ability participants, and suggest that *Digit Span Forward test*, serves as a “warm-up” task for high-ability participants. The *Letter Number Sequencing* task required participants to repeat a scrambled sequence of numbers and letters in a specific order: numbers first from the smallest to the biggest, followed by the letters in the alphabetic order. This task does not only measure the capacity of WM but also participant’s ability to extract and process the information store in WM. Compared with digit span tasks, it is believed to capture more subtle differences in WM associated with the central executive function (Crowe, 2000). *Arithmetic* required participants to process a series of mathematical questions under time limit without access to pen-paper calculation. This task is believed to assess the recall of both the question heard and previously learned mathematical rules. It is also considered to capture some of the attention aspects. Although it might be hard to control for participants’ mathematical background (e.g. whether some students but not others may have taken math modules), I nevertheless include this task as there are some recent study on bilinguals children suggesting a relationship between second language ability and arithmetic skills (Van Rinsveld, Brunner, Landerl, Schiltz & Ugen, 2015; Van Rinsveld, Schiltz, Brunner, Landerl & Ugen, 2016).

For attention measures, the current study used the Test of Everyday Attention (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994). This is a test which measures multiple aspects of both visual and auditory attention. Although the research focus is on auditory language, the language training task involved both visual and auditory stimuli (Mandarin words and tone diacritics, see Section 4.3.2) so that visual attention is relevant. Auditory and visual attention were assessed using six subtests (dropping those tasks that are only sensitive to individuals with attentional deficits, on which the participants are expected to be at ceiling):

*Elevator Counting with Distraction*, *Elevator Counting with Reversal*, *Visual Elevator*, *Telephone Search* and *Telephone Search while Counting*. The goal was to select a range of tasks that measure aspects of attention beyond just attention switching ability. *Elevator Counting with Distraction* required participants to count the low tones while ignoring the high tones. *Elevator Counting with Reversal* asked participants to count the middle pitch tone but change the counting direction to forwards or backwards upon hearing higher or lower tone. These two tasks are believed to reflect one's auditory-verbal attention and maybe particularly relevant to tone learning. While *Elevator Counting with Distraction* relates mainly to sustained attention and selective attention, *Elevator Counting with Reversal* assesses the ability to switch attention between auditory stimuli. *Visual Elevator* required participants to count a sequence of elevator signs out loud as quickly as possible and change the direction of counting upon encountering up or down arrows. This task measures visual attention and attention switching skills between auditory and visual stimuli. Finally, *Telephone Search* and *Telephone Search while Counting* asked participants to search for certain symbols next to telephone numbers on a yellow page, which assesses selective attention. *Telephone Search while Counting* additionally required participants simultaneously count tones and captures attentional switching and selective attention.

There is no consensus as to the best way in which to measure musical ability. Many studies used self-reported questionnaires (e.g. Swaminathan, & Gopinath) to try to quantify individual musical experiences such as the number of years spent learning instruments. However, such measures do not capture the core abilities involved in learning tonal languages, i.e. the ability to process pitch patterns. Another potential problem is that previous studies mainly used participants who were either proficient musicians (e.g. Delogu, et al., 2010) or people had received formal music training (e.g. Milovanov, et al., 2008). There is a lack of

studies exploring whether musical ability in the general population would affect language learning. The current study used the Goldsmiths Musical Sophistication Index (Gold-MSI). This is the test battery designed especially to assess the music sophistication of general population (Müllensiefen, Gingras, Musil & Stewart, 2014). Although this test-battery was developed relatively recently, it has been normed with a sample pool of more than 100,000 participants and has been used in a number of studies (e.g. for a meta-analysis on the relationship between musical ability using various measures and WM, see Talamini, Altoè, Carretti & Grassi, 2017). The current design took the tasks *Beat Perception* and *Melody Memory* from this battery. *Beat Perception* is a variant of the original beat alignment task developed by Iverson and Patel (2008). In this task, a music excerpt was played to participants and a series of beep sounds would also be played; the participants' task was to decide whether the beep sound was on or off the beat of the music. This task was believed to capture participant's ability to attend to the rhythmic feature of the music. This ability might seem to be most relevant when learning connected speech rather than monosyllabic single words, however, Mandarin tones differ in duration. At least when monosyllabic words are produced in isolation (as in the current study), there is agreement that tone 3 has the longest duration while tone 4 has the shortest duration, with tone 1 and tone 2 falling in between (Whalen & Xu, 1992; Yang, Zhang, Li & Xu, 2017). It is thus interesting to investigate whether the ability to attend to time-based rhythmic feature affects tone learning. The *Melody Memory* test required participants to listen to two pieces of music in different keys and decide whether they had an identical pitch interval structure. This task is similar to the Perceptive Tonal Memory Test from Wing Measures of Musical Talents (Wing (1968) used in the study by Li & Dekeyser (2017) discussed above) and assessed both participants' auditory WM as well as sensitivity to pitch patterns.

#### 4.2.2 *Training and testing paradigm*

In addition to including this battery of individual different tasks, several changes were made to the current training design compared with Study 1 and 2. Firstly, I decided to match the training paradigm to one that has been more generally used in the phonetic training literature, where participant learns to recognise a form of orthography which depicts some aspects of the phonological categories being learned (e.g. Bradlow, Akahane-Yamada, Pisoni & Tohkura, 1999). Since the characters used in Chinese orthography do not represent tones, in the current design I used Mandarin pinyin with tone diacritics. The motivation here was that it is possible that high variability might be particularly useful when using an identification task where there is a more direct representation of the category being learned (here Mandarin diacritics represent tone categories). Using this type of representation also allows me to have untrained items at pre-test as well as post-test, which is impossible in a design mapping novel words to pictures.

Another difference was that a 4AFC rather than 2AFC training method was used. In Study 1 & 2, for each of the Mandarin syllables heard in training, it was only used in one tone contrast (e.g. the base syllable [bi] was only used in tone 2 and tone 3). This was necessary due to the choice in that study to use a picture training paradigm where it is difficult to find syllables where combining with each of the four tones results in an imageable word. In contrast, in the current study, the choice to use pinyin representations for words meant that for each base syllable used, it is possible to present it with all four tones during training so that participants can make a more straightforward comparison between all four Mandarin tones. In order to maintain the same number of words in training as in the previous study, only half of the number of syllables from the previous study were selected. All of the words were real words in Mandarin. The training contains only four sessions, rather than six as in the previous studies,

since previous training data showed very little improvement in performance in the last two sessions.

For performance measures, as in Study 1 and Study 2 both perception and production tasks were included. For perception, as noted above, one of the tests was a pre-to-post version of the *Pitch Contour Perception Test* used in Study 1 & 2. Recall that the *Pitch Contour Perception Test* is a direct measure of participants' ability to identify Mandarin tones using the diacritics, and there was a significant pre- to post- training improvement on this task in previous studies. This may be even more likely here given that the new paradigm trains participants to identify the diacritics. I also include a test of categorical discrimination similar to the Three Interval Oddity task used in the previous study, but with an increase in the number of the items presented in each trial to make it a Four Interval Oddity task. The purpose of the change was to remove the difference in difficulty between trials (i.e. previously, it was found that participants performed better in trials where the target word was spoken by a speaker whose gender differed from the other two speakers, see section 2.3.4.1 & 3.3.4.1, the effect of *trial-type*); with four speakers, the task can balance the genders in every trial. Another difference was that in the current study, only untrained items were included in the test. This change was made since generalisation to novel items is of key interest and this design allows more items testing generalisation, increasing the overall power. For production measure, as the current design used Mandarin orthography, it was possible to use a reading task similar to the one used in previous literature (Li & DeKeyser, 2017) and it can be used at both pre- and post-tests with trained and untrained items. Similar to Study 1 and 2, both tone and Pinyin accuracy will be measured.

An important aspect of the current design is that since all of the tests are conducted pre- and post- training, for all of the individual difference measures it is possible to look at both

whether this predicts baseline performance in the test (*ID measures* predicts pre-test performance in the test) *and* correlations with the change due to training (*ID measure* predicts effect of *test-session*). In each case, there is a positive predictive relationship (allowing for a one-tailed test). Since there are two participant groups, in each case, for both whether the *ID measures* predicts performance and whether it predicts pre-to-post improvement, I also investigated whether this is modulated by participant's condition. Here there was no clear directional prediction as to which group will show a larger effect so a two-tailed hypothesis is tested. It is also possible that current learners of a tone language might begin the experiment with a different cognitive profile than those not currently studying a language. Thus the current study also explored whether at pre-test, there was a difference between the two participant groups in terms of their performance on each *ID measure*. As in Study 2, BF statistics was conducted in order to be able to evaluate evidence for the null. However in this study, Bayes Factors were used from the outset as the primary method of inference for the analyses with the individual difference measures (where in Study 1 & 2 they were used only when a null results was previously found with frequentist statistics).

In sum, the current study aimed to explore further how cognitive individual differences affect the efficiency of high-variability training for Mandarin tones. This is first study to look at the effects of WM, Attention and Musical Ability altogether in high variability training of lexical tones. Participant' individual differences will be measured in terms of WM (WAIS-III), Attention (TEA), Musical Ability (Gold-MSI) and the ability to identify the Mandarin tones (*Pitch Contour Perception Test*, pre-test). They will receive a four-session training programme. Their learning will be tested using the pre-to-post tests: Four Interval Oddity (categorical discrimination), Pitch Contour Perception Test (identification) and Picture Naming (production).

### 4.3 Method

#### 4.3.1 Participants

Sixty adults participated in the experiment. Twenty native English speakers who had no previous knowledge of Mandarin or any other tonal language (naïve participants - NP) were recruited from UCL Psychology Subject Pool. Forty students of Mandarin (Mandarin learner participants - MLP) were recruited from SOAS BA Chinese programme. Participant information is summarised in *Table 15*. It can be seen that participants in the MLP group were younger than those in the naïve group and this was significant ( $t(58) = 8.10, p < .01$ ). 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.

For the naïve participants group, all participants were native speakers of English who had not been exposed to another languages during childhood. None had any prior experience of Mandarin Chinese or any other tonal language. They had learned other second languages (average number 1.8 ( $SD = 0.4$ )) at high school level and the average age for starting to learn the first L2 was 13.42 ( $SD = 1.85$ ). Eighteen of these participants were graduates from universities, while two were university students in their 3<sup>rd</sup> year. For the MLP group, they were all year two undergraduate students at SOAS who had studied Mandarin for 18 months. Under the SOAS regulation they have all reached HSK Chinese Proficiency Test level 2. The programme teaches basic Mandarin knowledge (e.g. pronunciation, grammar and Chinese characters) and Chinese history and literature from earliest times up to the present. MLPs were also native speakers of English who had not been exposed to another languages during childhood. Other than Mandarin, they had learned other second languages (average number 2.3 ( $SD = 0.2$ )) at high school level and the average age for starting to learn the first L2 was 14.01 ( $SD = 0.77$ ).

Table 15: 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	Age that began L2 learning
Naïve participants	24.3	19-29	1.8	12-15
Mandarin Learner participants	19.75	19-24	2.3	13-15

#### 4.3.2 Stimuli

These stimuli consisted of the 18 Mandarin syllables used in Study 1 & 2. In order to avoid accidental differences between trained and untrained words (as observed in the previous two studies), words were reassigned to be either trained or untrained items (training/tests) with a view to balancing the difficulty across these (based on how well they were learned at pre-test in previous studies, see Appendix B & C). In this study, trained words were only used in Training and Pinyin reading while untrained words were used only in Four Interval oddity. For all 18 trained words, all four tones were assigned to each syllable and I made sure each word was still a meaningful word in Mandarin (e.g. *māo*, Tone 1 [*cat*]; *máo*, Tone 2 [*fur*]; *mǎo*, Tone 3 [*wooden rivet*]; *mào*, Tone 4 [*hat*]), resulting in 72 words in total. For the other 18 untrained words, the same tone contrast patterns were used as in Study 1 & Study 2, such that they formed 18 minimal pairs of Mandarin words (3 minimal pairs for each of the six tone contrasts generated by the four Mandarin tones, see Appendix C).

The full set of 108 Mandarin words were recorded by two groups of native Mandarin speakers using a Sony PCM-M10 handheld digital audio recorder. The first group was made up of two newly recruited female speakers and two male speakers and they recorded all stimuli used in Training. The second group consisted of two female speakers and two male speakers used in the training in Study 1 & 2. These stimuli were used in the Four Interval Oddity task.

For the *Pitch Contour Perception Test*, the same set of stimuli was used as in Study 1 and Study 2, i.e. six Mandarin vowels (/a/, /o/, /e/, /i/, /u/, /y/) repeated in the four Mandarin tones by the two male and two female native Mandarin speakers who produced the stimuli for the Four Interval Oddity making 96 stimuli in total.

All words were edited into separate sound files, and peak amplitude was normalised using Audacity (Audacity team, 2015, <https://sourceforge.net/projects/audacity/>). Any background noise was also removed. All recordings were perceptually natural and highly distinguishable as judged by native Chinese speakers. Overall, all stimuli were identical across conditions and participants. The training paradigm and the tests used two different sets of speakers and items.

#### 4.3.3 Procedure

The experiment involved three stages (see Figure 25): Pre-test (session 1), training (sessions 2-5), and post-test (session 6). Participants were required to complete all six sessions within two weeks, with the constraint of one session per day at most. The pre- and post-test sessions took place in a quiet, soundproof testing room in Chandler House, UCL. The training sessions were undertaken by participants on their own machines at home using an online testing platform Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2018).

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. For both groups, 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. For the Pinyin reading task only, I used the same computer program as Study 1 & 2 (which was custom-built software package developed at the University of Rochester). All other tasks were carried out on Gorilla.

The specific instructions for each task were displayed on screen or explained to the participant 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 order in Figure 25 and described in more detail below. There was no time limit for making responses in any of the tasks. Participants wore a pair of HD 201 Sennheiser headphones in the lab based sessions and were also instructed to use them for the training sessions at home.

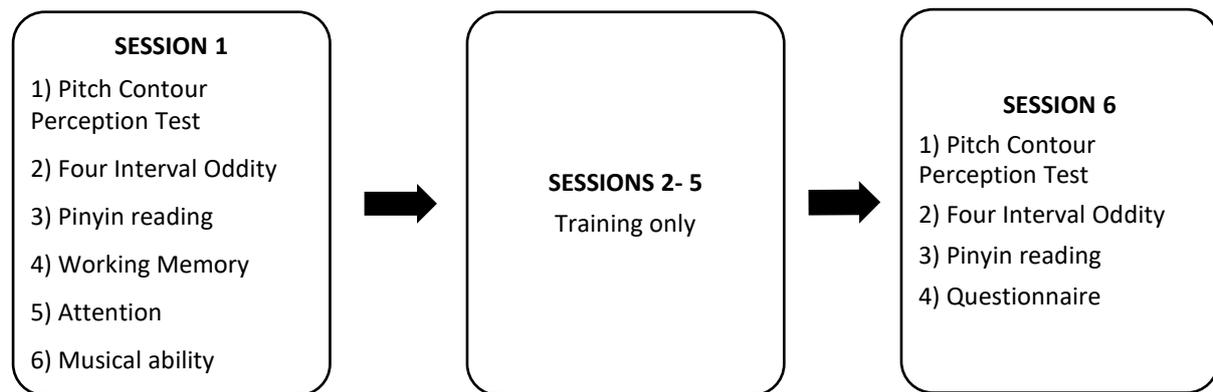


Figure 25 Tasks completed in each of the six sessions.<sup>9</sup>

#### 4.3.3.1 The Pitch Contour Perception Test

This test was identical to the equivalent test used in Study 1 and Study 2 (see Section 3.3.2). As in the previous study, I used pre-test performance in this task as a measure of individual aptitude. However, as discussed, since the nature of this task was identifying Mandarin tones- i.e. the same ability targeted during training- and since in the previous study the analysis *did* find improvements in this task after training, comparison of pre- and post-training performance was also used as an outcome measure.

---

<sup>9</sup> *ID measures* (working memory, attention & musical ability) were also measured again in post-test. However, I realised later that data from this task would be sensitive to practice effects and so I do not analyse and report this data.

#### 4.3.3.2 *Four Interval Oddity Test*

This test was similar to the Three Interval Oddity task used in Study 1 & Study 2 (see section 2.2.3.3) with one major difference: there are four items used in each trial, rather than three, with each item produced by a different speaker (2 females, 2 male) (thus avoiding the differences between trials depending on the balance of male/female speakers across trials exhibited in previous study). Using four tokens also means that the overall difficulty of the test increased, allowing it to avoid potential ceiling effects which might arise due to the recruitment of current Mandarin learners and the use of Mandarin Pinyin in training (see section 4.3.3.5). Each of the words in the minimal pair was used once as the target (“different”) word, making 72 trials in total.

#### 4.3.3.3 *Pinyin Reading Test*

The pinyin representations of each of the 72 words used in training which included tone diacritics were presented in a randomised order. Participants were instructed to try to pronounce the Mandarin word. Verbal responses were audio recorded and were later transcribed and rated by native Mandarin speakers (see section 4.4.2.2.1).

#### 4.3.3.4 *Measures of Working Memory, Attention and Musical ability*

Working memory scores were measured using WMI from WAIS-III test (Wechsler, 1997), specifically the *Digit Span (forward and backward)*, *Letter Number Sequencing*, and *Arithmetic* tasks. Auditory and visual attention were assessed using the six subtests (*Elevator Counting with Distraction*, *Elevator Counting with Reversal*, *Visual Elevator*, *Telephone Search* and *Telephone Search while Counting*) in the Test of Everyday Attention (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994). Musical Ability was measured using The Goldsmiths Musical Sophistication Index (Gold-MSI) (Müllensiefen, et al., 2015) using the *Beat perception* and *Melody Memory* tasks. All of these were administered following the respective instruction manuals.

#### 4.3.3.5 Training Task

Participants completed the training task in Session 2-5. The task was similar to Study 1 and Study 2, however here in each trial, participants heard a Mandarin word and then had to select one of *four* (rather than two) candidate responses, and here the responses were pinyin representations of the words including a diacritic representation of the tone. All four words were displayed on the computer screen. The four word choices always had the same pinyin spelling but with each of the four diacritics (see Figure 26). After selecting a word, the participant was informed whether their answer was correct or incorrect. If the correct choice was made: (i) the participant was awarded a coin and the “coin number” shown in the lower right corner increased by one (ii) the participant saw the correct Pinyin (the other three vanished from the screen) and heard the correct word again. If the participant answered the trial incorrectly: (i) the participant saw the incorrect Pinyin they had chosen on a new screen (the other three words disappeared) and heard the *corresponding* word (i.e. the word which actually matched the pinyin displayed on the screen) (ii) the participants then heard the target word again whilst viewing the target pinyin. This was motivated by the work of Iverson & Evans (2009), who presented both the incorrect item chosen by the participant and the correct item to choose during HVPT. It is considered to be able to facilitate the categorical discrimination between the stimuli.

There were 72 words used. Each word was used as the target word four times, each time pronounced by a different speaker, resulting in 288 trials in total per session. In each session, participants heard all four speakers, each in a separate block but with the order of the items in each block randomised. Each training session lasted for approximately 30 minutes and the number of coins earned was displayed at the end of the session.

Overall, this task increased the number of pictures displayed in each trial which increased the difficulty compared to Study 1 and Study 2. However, the use of Mandarin Pinyin

and tone symbols should ease the burden on memory, thus reducing the difficulty of the task, and so that I expected the training paradigm to remain effective.

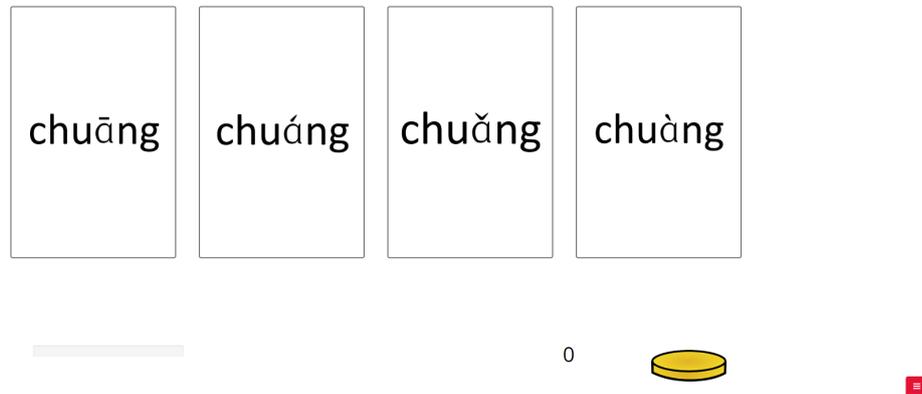


Figure 26 Screen shot from the training task. The stimuli heard is ‘chuang’, tone 4.

#### 4.3.3.6 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. This information is reported in

*Table 15* and was also used to check that they did not know any other tonal language and they were not childhood bilinguals (here defined as having learned second language before the age of 5; for a discussion regarding the onset of bilingual acquisition, see Schulz & Grimm, 2018).

## 4.4 Results

#### 4.4.1 Comparison of the participant groups for each of the individual aptitude tasks

These analyses examine whether there is a difference between the naïve participants (NP) and Mandarin Learner Participants (MLP) groups on any of the *ID measures* which are later used as predictors in the analyses reported in Section 4.4.2. Table 16 shows the standardised scaled scores of working memory and attention measures. It can be seen both the NP and the MLP perform approximately within the normal range (i.e. within 7-13 – one SD above/below the mean), although for the MLP group, their average performance on *Digit Span Forward* (13.88) and *Visual Elevator* (13.24) was slightly above. For the music measures, *Beat Perception* and *Melody Memory*, no standardised score is available, however all of the participants performed within the 5th to 95th percentile according to the Gold-MSI manual. Figure 27 (Working Memory), Figure 28 (Attention) and Figure 29 (Musical Ability) show the distributions of participants' raw data<sup>10</sup>. It should be noted that in some cases (*Elevator Counting with Distraction*, *Elevator Counting with Reversal*, *Beat perception* and *Melody Memory*) here and in the analyses, the raw scores were multiplied by 10 to make the calculation process easier. From the figures it can be seen that in general, the MLP group showed better performance (greater accuracy/ faster responses) than the NP group.

---

<sup>10</sup> Please note that although pre-test *Pitch Contour Perception Test* score is used as an ID measure, I do not report comparisons for the two participant groups here, since this is included in the analysis reported in section 4.4.2.4.

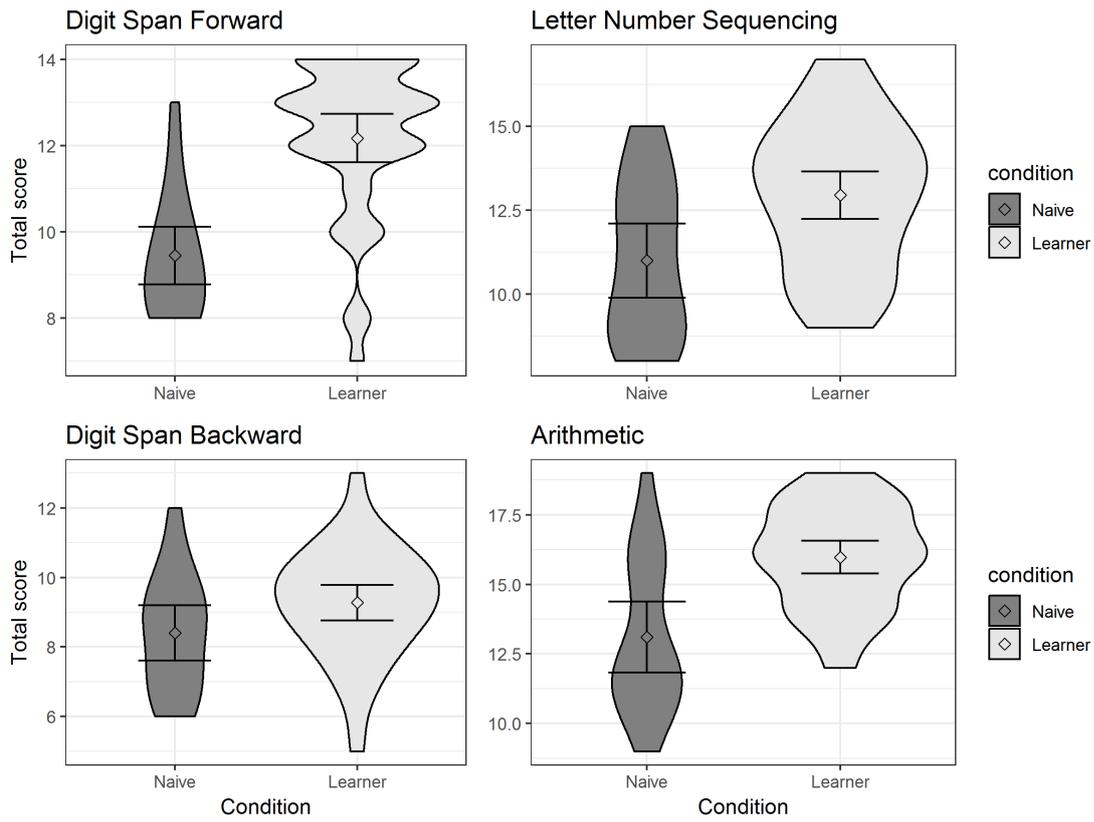


Figure 27 Mean proportion of correct of Digit Span Forward, Digit Span Backward, Letter Number Sequencing and Arithmetic for naïve participants and Mandarin learners.

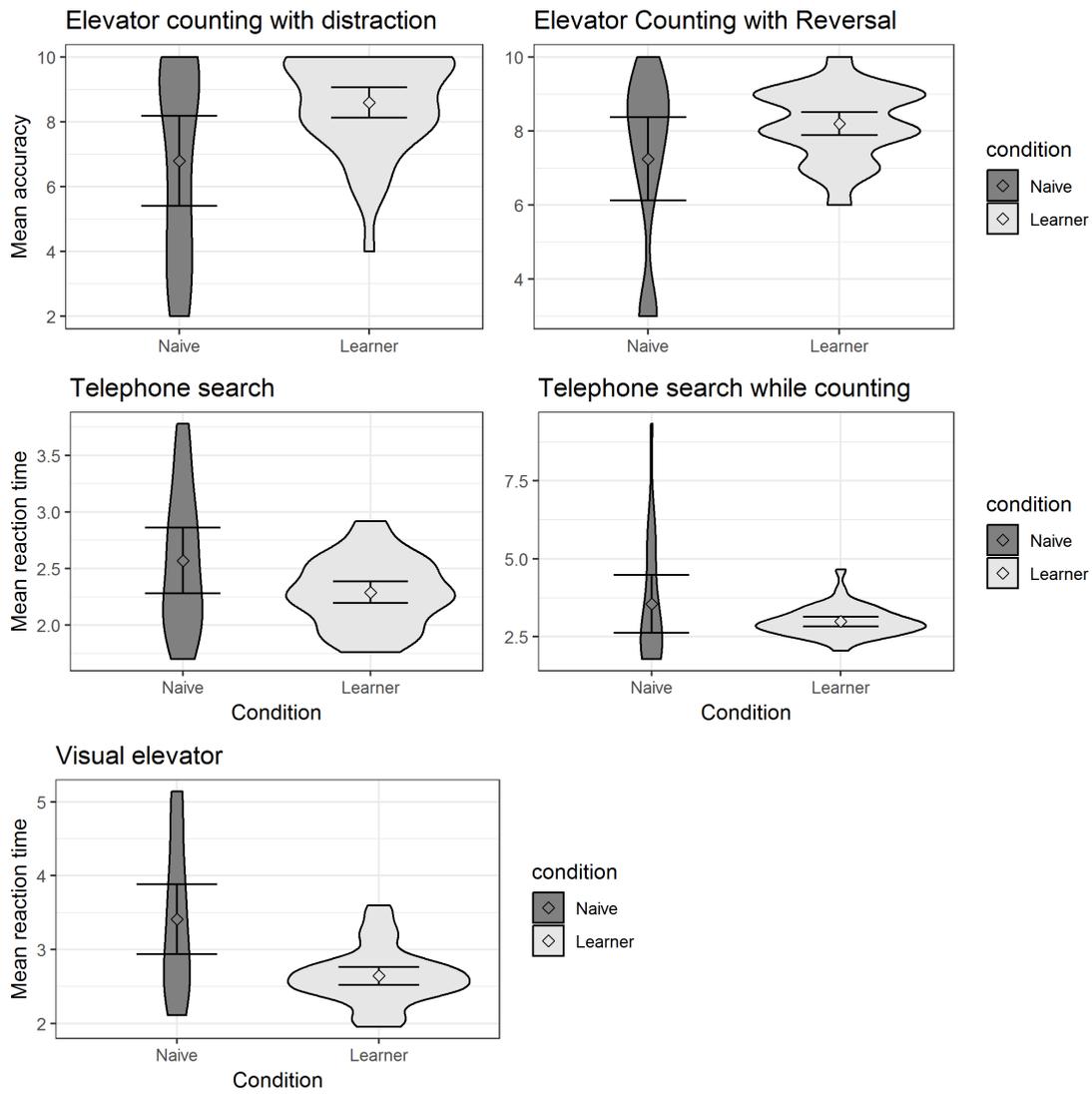


Figure 28 Results of Elevator counting with Distraction, Elevator Counting with Reversal, Telephone Search, Telephone Search while Counting and Visual Elevator for naïve participants and Mandarin learners.

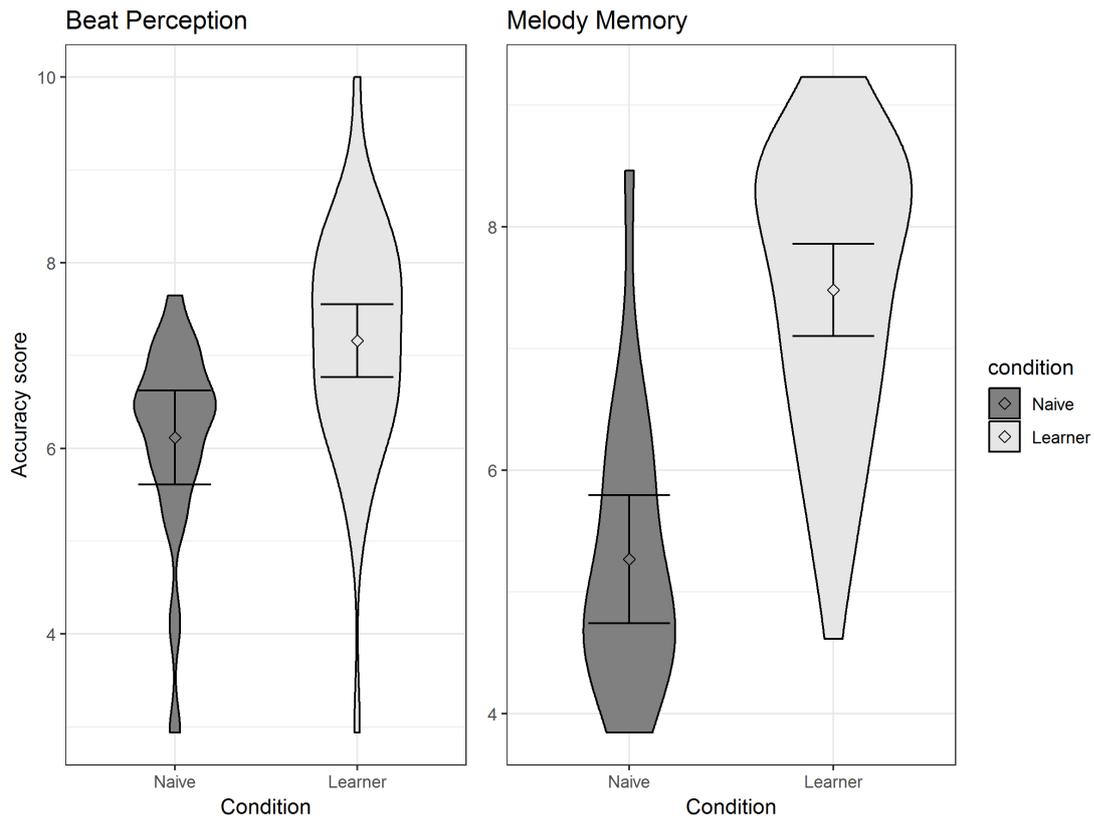


Figure 29 Mean proportion of correct of Beat Perception and Melody Memory for naïve participants and Mandarin learners.

Table 16 Standardised scores ( $M = 10$ ,  $SD = 3$ ) for each individual difference measure at pre-test. Numbers in brackets are standard deviations.

Task	Learner		Naive	
Digit Forward	13.88	(2.33)	12.57	(2.09)
Digit Backward	11.85	(2.58)	11.37	(2.64)
Letter Number Sequencing	11.95	(2.21)	11.3	(2.36)
Arithmetic	12.48	(1.54)	11.88	(2.39)
Elevator Counting with Distraction	10.38	(2.24)	9.82	(3.10)
Visual Elevator	13.24	(2.06)	12.13	(3.99)

Elevator Counting with Reversal	11.2	(0.97)	10.82	(2.78)
Telephone Search	12.38	(2.87)	11.82	(4.04)
Telephone Search while Counting	8.94	(0.69)	8.82	(3.25)

---

In order to further examine the difference between the two groups, a series of linear models were run with the raw scores for each of the *ID measures* as the DV. In total, 12 models were run and the results are displayed in *Table 17*.

Here both *p*-values and Bayes Factors are reported. For *p* values, note that the reported values are *not* adjusted for multiple hypotheses testing, under the view that these tests are not *independent* tests of the same hypothesis (e.g. *Digit Span forward, Digit Span Backward, Letter Number Sequencing & Arithmetic* are working memory measures which are likely to be correlated to some extent (e.g. Ryan & Paolo, 2001) - see also the PCA analysis in Chapter 4.6). However, given the large number of tests (11) carried out at the same time, these values should be treated with caution. Bayes Factors were also computed as a measure of strength of evidence. Multiple correction was also not used for these, as is standard for Bayes Factors which remain a valid measure of the evidence regardless of how many hypotheses are tested. Significance testing depends on computing probabilities and given the same  $\alpha$  level, by definition the probability of finding a false null result increases as we increase the number of tests. On the other hand, correcting for multiple hypothesis testing introduces inferential arbitrariness. Instead, Bayes Factors (where H1 is motivated by theory) involve relating theories to data and contrasting two models (for H1 and H0). A Bayes Factor therefore gives a continuous measure of evidence for each model. This tells what the data indicate, and this measure of evidence is valid regardless of what other theories are tested (for a detailed

discussion, see Dienes, 2016). I took the same approach as in Study 1 and 2: the main data was the betas and SEs from the models and I modelled H1 as a half-normal distribution (since I expect the MLP, who are current university students, to outperform the NP group) with a mean of 0 and an SD of  $x$  where  $x$  is an estimation of the predicted difference. In the absence of any prior data using sufficiently similar materials, and since I did not wish to use unprincipled default values, I estimated  $x$  as follows.: For the working memory and attention scores given the participants are all typically developed, I assumed the greatest possible difference between the two groups in scaled scores would be 1 SD above and 1 SD below the mean. Therefore, I checked the manuals (WAIS-III & TEA) to find the corresponding raw scores for 1SD above 1SD below the mean for each task. I modelled H1 using half of the distance between these raw scores (since I am estimating the SD of a half-normal distribution, so the maximum should be approximate equal to  $2x$ ). For Musical Ability measures (where there are no scaled scores) I computed a maximum difference as the difference between the maximum and the minimum raw score and thus set  $x$  to half of this difference.

As in Studies 1 and 2, I interpret BFs using the following conventions:  $B < 1/3$  indicates substantial evidence for the null,  $B > 3$  indicates substantial evidence for H1, values between  $1/3$  and  $3$  indicate that the data collected do not sensitively distinguish H0 from H1 (Jeffreys 1961; Dienes 2008). Since there is subjectivity in how the values for H1 are determined, I indicate the robustness of Bayesian conclusions by reporting a robustness region for each  $B$ , which gives the range of values of the scale factor  $x$  that qualitatively support the same conclusion (i.e. evidence as supporting H0, or as supporting H1, or there not being much evidence at all).<sup>11</sup> The ranges and corresponding Bayesian analyses are summarised in Table

---

<sup>11</sup> To find out about the range of values, in each case I started at 0 (i.e. no difference between conditions) and went through 100 equal steps up to a value *max*; *max* was what I considered to be the largest possible difference between the two conditions given the scale. To determine the value *max*, for tasks with accuracy score measures,

17. It can be seen that the MLP group performed better than the NP group on all *ID measures* with Bayesian analyses showing strong to substantial evidence, with the exception of the attention measure TS measure, where the evidence was ambiguous.

*Table 17* Regression results and Bayesian factors for individual difference measures at pre-test. Positive  $\beta$  indicates larger effect in the MLP group, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results.

Task	$\beta$	SE	p	Length of predictor <sup>12</sup>	Robustness range	H1	Bayes	Robustness Region
Age	-4.55	0.49	<0.001	11				
Digit Span - Forward	2.73	0.46	<0.001	7	[0, 16]	0.52	31864.90	[0.16,>16]
Digit Span - Backward	0.88	0.44	0.056	8	[0, 14]	0.602	3.864	[0.42, 1.56]
Arithmetic	2.88	0.60	<0.001	10	[0, 22]	0.56	340.85	[0.44,>22]
Letter - Number Sequencing	1.95	0.62	0.003	9	[0, 21]	0.867	30.932	[0.64,17.39]
Elevator Counting with Distraction	1.8	0.57	0.003	8	[0, 10]	0.896	37.775	[0.51,>10]
Elevator Counting with Reversal	0.95	0.43	0.03	7	[0, 10]	0.675	5.65	[0.30, 2.83]
Visual Elevator	-0.77	0.18	<0.001	3.19	[0, 7.8]	-0.6	2432.63	[0.16,>-7.8]
Telephone Search	-0.28	0.12	0.021	2.08	[0, 6.4]	-0.5	6.54	[0.13,-1.23]
Telephone Search while Counting	-0.57	0.32	0.09	7.56	[0, 12.4]	-1.5	1.733	[0,-8.52]
Beat Perception	1.04	0.32	0.002	7.06	[0, 10]	3.529	31.446	[0.30,>10]
Melody Memory	2.21	0.32	<0.001	5.38	[0, 10]	2.692	>99999	[0.10,>10]

this is equivalent to the number of items in the task (for a task with 24, items, 24 is the max). For Tasks with RT measures, I took this to be the difference between the slowest response time registered on the TEA manual and 200ms (generally regarded to be fastest possible response time e.g. for filtering). In some cases I did not find the end of the robustness region within this range, in this case: I denoted the end of the range as “,> max”. In addition, for BF <1/3, the end of the robustness regions is always infinity, as written in  $\infty$ .

<sup>12</sup> Length represents the difference between the maximum and minimum score for that *ID measure* in this dataset. This is used in estimating the predicted value of H1 used in the Bayes factor calculations, as detailed in the section below.

#### 4.4.2 Performances measures: Tests Administered Pre- and Post- Training

##### 4.4.2.1 General statistical approach

As for Study 1 and 2, analyses for each of the performance measures –here Pinyin Reading (Pinyin accuracy and Tone accuracy), Four Interval Oddity and Pitch Contour Perception Test are reported separately. In each case, first a model without any individual difference measures was run. Performance in the relevant task was the DV and logistic mixed effect models were used (since the DV was binary in each case). There were two factors: *participant-condition* with two levels (Naïve participants (NP) /Mandarin learner participants (MLP), which was given a numeric centered coding, and *test-session* with two levels (pre-test/post-test) which was coded as a factor with “pre-test” set as the reference level- allowing us to look at the possible differences between the experimental conditions at the pre-test stage, as well as whether post-test performance differed from this baseline. In all analyses I automatically put experimentally manipulated variables and all of their interactions into the model (i.e. without using model selection). The questions of interest were: (i) whether the two groups differed on this task at pre-test (effect of condition at pre-test) (ii) whether participants improved on this test following training (effect of session) and (iii) whether the possible improvement observed in (ii) differed between groups (effect of session by condition).

Following these analyses, an additional set of analyses was conducted for each of the outcome measures looking at the role of individual difference in predicting performance. I have 11 *ID measures* from the standardised tests for each participant taken at pre-test (4 Working Memory measures from WAIS-III, 5 Attention measures from TEA and 2 Musical Ability measures from Gold-MSI). In each case, participants’ mean score on the *ID measure* is used as a predictor. Their mean *Pitch Contour Perception Test* score at pre-test was used as an additional *ID measure* predictor in Four Interval Oddity and Pinyin Reading tasks. Given the

difference between the age groups reported in the section 4.3.1, age was also included as a predictor (with the hypothesis that there will be a negative relationship, whereby younger is better, to see if there is evidence of any benefit for the younger participants). Additionally, similar to Ou et al. (2015) who computed composite scores for Working Memory and Attention, I also computed three *composite scores* for Working Memory, Attention and Musical Ability aiming to acquire more robust general measures of these abilities. This was done by transforming the participants *ID measures* into z scores (i.e.  $M = 0$ ,  $SD = 1$ ) and summing them (for *Visual Elevator*, *Telephone Search* and *Telephone Search while Counting* where smaller RT measures represented better performance, the sign was changed before computing the composite score).

Each of these 16 *ID measures* was added as a predictor into a separate version of the model for each outcome measures (except that *Pitch Contour Perception Test* was not used in the model with *Pitch Contour Perception Test* score as the outcome measures) along with the interactions: with *ID-measure by condition*, *ID-measure by test-session* and *ID-measure by test-session by* and *participant-condition* making the following number of additional models for each of the original four models : 15 (*Pitch Contour Perception Test*); 16 (*Pinyin reading-tone accuracy*, *Pinyin reading-Pinyin accuracy*, *Four interval Oddity*). This allowed me to examine: (i) whether the *ID measures* predicted participants' performance on this task at pre-test (effect of *ID measure* at pre-test) (ii) whether the effect observed in (i) differed across different *participant-condition* (*ID measure* x *participant-condition* interaction) and (iii) whether the effect observed in (i) differed across different *test-session* (*ID measure* x *test-session* interaction) (iv) whether there is any three-way interaction, that the effect of individual aptitude measures differed across both *test-session & participant-condition* (*ID measure* x *test-session* x *participant-condition* interaction). As in the previous section, I reported *p* values

without adjustment for multiple comparisons, though note that these should be treated with caution. Bayes Factors were also provided as a measure of the evidence for the hypotheses and I use these as the key method of inference. As discussed in Section 4.4.1, Bayes factors remain a valid measure of evidence regardless of the number of hypotheses tested. However, Dienes (2016) discusses the approach of meta-analytically combining measures and looks for evidence for these combined measures. In the current analysis, I use similar composite measures to test more general theories (e.g. working memory is generally predictive of performance) alongside more specific measures (e.g. performance on a digit span backwards task is predictive of performance). Dienes (2016) discussed this approach of explicitly relating data to more/less specific theories using Bayes Factors and suggested it as a more principled, approach than arbitrary corrections for multiple hypothesis testing.

To compute Bayes Factors: I again took the betas and SEs for the relevant coefficients from the linear/logistic models as reported in each section (note that for binary measures, this allows to meet normality assumptions by continuing to work within log-odds space). I modelled H1 as either a half-normal distribution or a normal distribution. Half-normal distribution was used where I had a clear directional prediction, i.e. that the *ID measure* should positively predict participants' performance at pre-test and that they should improve from pre-to post-test. However, I did not have a clear directional prediction regarding whether the groups will differ either in the extent to which performance is affected by the *ID measure* at pre-test, or in the extent that *ID measure* predicts pre-to-post improvements. Thus, for these hypotheses I used a normal distribution. In both cases, I assume the (half-) normal distribution has a mean of 0 and an SD of  $x$ . In the absence of any prior data using sufficiently similar materials, and since I did not wish to use unprincipled default values, I estimated  $x$  for each analyses at follows:

*Hypothesis that the ID measure predicts performance at pre-test:* I set  $x$  as the difference between the grand mean at pre-test (the Intercept - since I coded with pre-test as the reference level) and an estimate of baseline performance on the task, divided by the length of *ID measure* (i.e. by the difference between the maximum and minimum score for that *ID measure* in this dataset,  $t$ ). The logic is as follows: the maximum effect of the ID predictor on the experimental outcome is seen if participants with the lowest value of the *ID measure* are at baseline in the experimental task, and performance in that task changes linearly with each step of the *ID measure*. If performance on this test is  $p$  (so the grand mean is  $\bar{p}$ ), the baseline is  $b$ , and the predictor has  $n$  levels, the effect of a one-step change in the predictor on the outcome will be equal to:  $2(\bar{p}-b)/n$ <sup>13</sup>. This gives us an estimate of the *maximum* value of  $x$ ; since I am using a half normal distribution with a mean of zero, I assume the maximum value is equal to approximately  $2SD$ , and I can set the estimate  $x$  of the standard deviation to be equal to *half* of this value (i.e.  $x = (\bar{p}-b)/n$ ). The estimate of the baseline  $b$  depends on the task: for the tone measure of Pinyin Reading test, I assume a 1/4 chance that the rater could identify the correct tone out of the four possibilities one (25% = -1.099 in log odds space). For the pinyin measure of the Pinyin Reading test, since there is no clear chance value, I estimated “baseline” performance on the basis of what might be expected for literate English speakers reading the script as though it were English orthography. Specifically, for those words which shared the same phonetic pronunciation in English and Mandarin, I assume correct pronunciation; for those words that shared similar phonetic pronunciations (e.g. dao), I assume

---

<sup>13</sup> The logic is as follows: if a one-step change in the predictor is equal to  $s$  then,  $\bar{p} - b = \frac{0+1(s)+2(s)+\dots+n(s)}{n+1}$ . Applying the formula for triangular numbers:  $\bar{p} - b = \frac{s(n)(n+1)}{n+1}$ . Rearranging the formula:  $(\bar{p} - b)(n + 1) = \frac{s(n)(n+1)}{2}$ ,  $2(\bar{p} - b)(n + 1) = s(n)(n + 1)$ ,  $2(\bar{p} - b) = s(n)$ ,  $s = 2(\bar{p} - b)/n$

there is a 50% chance<sup>14</sup> that the participants can produce the word such that it is judged as “correct” by the rater, which can be seen in Appendix B. On this basis, I compute an 8/18 performance as baseline in this task ( $8/18\% = -0.223$  in log odds space); for Four Interval Oddity Task and for PCPT, I assume a 1/4 chance of identifying the correct one ( $25\% = -1.099$  in log odds space).

*Hypothesis that the participant-conditions differ in the extent that the ID measure predicts performance at pre-test:* I set  $x$  as equal to the mean difference between NP and MLP (i.e. the beta for the main effect of *ID measure* at pre-test). The logic is as follows: the *maximum* difference is seen if one of the participant groups shows no effect of *ID measure* and the other shows a positive effect of *ID measure* (or a negative effect in the case of the RT measures). In this case, if the mean effect of *ID measure* is  $\bar{I}$ , the difference in  $I$  between the two conditions will be equal to  $2\bar{I}$ . Again, I can set the estimate of  $x$  to be half this value (i.e.  $x = \bar{I}$ ). Note that if there is no evidence of an overall effect in the predicted direction *ID measure* at pre-test (i.e. the main effect at pre-test is negative, or positive in the case of the RT measures), then I can’t use this method to estimate H1. In this case, I don’t compute a Bayes Factor for this interaction.

*Hypothesis that ID measures predict pre to post-test improvement:* I set  $x$  as half of the difference between the maximum possible effect of the *ID measure* on post-test performance and the actual effect of aptitude at pre-test (main effect of *ID measure*). The logic is that this half of the maximal difference in the effect of ID from pre to post-test. The maximum aptitude is defined as the difference between ceiling performance possible on the test given the scale,

---

<sup>14</sup> This estimation of 50% chance of producing the “Mandarin-like” sounds is somewhat arbitrary (and there is no similar estimation in the literature). The goal here is to have a rough estimate which can be used to compute a rough estimate of effect size ( $x$ ). In general, there is always some subjectivity in the choice of values of  $x$  used to inform H1 for Bayes Factor analyses. It is for this reason that I always include robustness regions which give the range of values of  $x$  that qualitatively support the same conclusion.

and the baseline performance in that task (as described above), divided by *ID measure* length. For the calculation of ceiling performance for the tone measure and pinyin measures of Pinyin reading and Four Interval Oddity, I estimated this as one incorrect 71/72 (4.263 in log odds space); for PCPT, I estimated this as one incorrect i.e. 95/96 (4.554 in log odds space). (Recall that I cannot compute log odds for the true ceiling of 100% in each test).

*Hypothesis that ID measure predicts pre to post-test improvement more for one participant group than the other:* I set  $x$  as equal to the mean difference between the pre-post improvement of the NP and the MLP groups (i.e. the beta for the interaction of *ID measure* and *test-session*). The logic is as follows: the *maximum* difference is seen if one of the participant groups shows no effect of *ID measure* x *test-session* and the other shows a positive effect of *ID measure* x *test-session* (or a negative effect in the case of the RT measures and age). In this case, if the mean effect of *ID measure* x *test-session* is  $\bar{I}$ , the difference in  $I$  between the two conditions will be equal to  $2\bar{I}$ . Again, I can set the estimate of  $x$  to be half this value (i.e.  $x = \bar{I}$ ). Note that if there is no evidence of an effect in the predicted direction of *ID measure* from pre- to post- test (i.e. the beta of *ID measure* x *test-session* interaction is negative, or positive in the case of the RT measures), then I can't use this method to estimate H1. In this case, I don't compute a Bayes Factor for this interaction.

Wherever there is evidence for a difference between the *participant-conditions* (either in the effect of ID measure at pre-test, or for the effect of ID measure on pre- to post- test improvement), I run separate models for the NP group and the MLP groups. In this case, what is of interest is determining whether the effect found actually holds for both groups. For example, an *ID measure* x *test-session* interaction suggests the extent to which *ID measure* predicts improvement from pre- to post- test differs by *participant-condition*. I therefore

investigate whether each group separately show an interaction between ID measure and pre- to post-improvement. To compute  $x$  in each case, where possible, I use the effect in one group to inform H1 when looking at the other group which did not show an effect (e.g. if the MLP group show an effect of *ID measure* by *test-session*, the beta for this effect in the MLP group can be used to estimate the same effect for the NP group, and so I set the SD of H1 –  $x$ - to this value). Where this isn't possible, I used the procedure laid out above (e.g. when looking at the effect of an *ID measure* at pre-test for the MLP group, if the NP group did not actually show this effect, I set  $x$  as the difference between the grand mean at pre-test and an estimate of baseline performance on the task, divided by the by the length of *ID measure*).

I again interpret BFs using the following conventions:  $B < 1/3$  indicates substantial evidence for the null,  $B > 3$  indicates substantial evidence for H1, values between 1/3 and 3 indicate ambiguous evidence (Jeffreys 1961; Dienes 2008) and again I also report robustness regions for each  $B$ , which gives the range of values of the scale factor  $x$  that qualitatively support the same conclusion<sup>15</sup>). The full set of analysis performed can be found at [https://osf.io/j6s7w/?view\\_only=497e0e8ee7ff4e7387984690eafd4b5a](https://osf.io/j6s7w/?view_only=497e0e8ee7ff4e7387984690eafd4b5a)

#### 4.4.2.2 Pinyin Reading test

##### 4.4.2.2.1 Coding and inter-rater reliability analyses

All stimuli were rated by Rater 1, who was recruited from the UCL MA Linguistics program and was naïve to the purposes of the experiment. Rater 2 was myself and rated 10% of the trials. Raters were presented with recordings in blocks in a random sequence (blind to

---

<sup>15</sup> To find out about the range of values, in each case I started at 0 (i.e. no difference between conditions) and went through 100 equal steps up to a value *max*; *max* was what I considered to be the largest possible difference between the two conditions given the scale. Since all of the outcome measures are binary, I set the value of *max* to 5, equivalent to 99% accuracy in log odds space (recall I can't reach 100% with binary measures). In some cases I did not find the end of the robustness region within this range, in this case: I denoted the end of the range as “,> *max*”. In addition, for BF <1/3, the end of the robustness regions is always infinity, as written in  $\infty$ .

*test-session* and *participant-condition*). For each item, raters were asked to identify the tone, and transcribe the pinyin (segmental pronunciation) produced by the participants.

Two measurements were taken from the production tasks: mean accuracy of tone identification (Tone accuracy- scored as 1/0) and 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)). The inter-rater reliability was examined for both measures using kappa statistics were calculated by the “fmsb” package in R (Cohen, 2014). Tone accuracy  $kappa = 0.56$  (“moderate agreement”), and for Pinyin accuracy  $kappa = 0.66$  (“moderate agreement”; Landis & Koch, 1977). All of the analyses presented in Section 4.4.2.2 were based on the ratings of Rater 1 (the naive rater).

#### 4.4.2.2.2 Tone accuracy

##### 4.4.2.2.2.1 Analysis of performance (without *ID measures*)

The predicted variable was whether the correct response was given (1/0) on each trial. The predictors were *test-session* and (pre-test, post-test) *participant-condition* (naïve, learner). The mean accuracy is displayed in Figure 30. Overall, participants performed better after training ( $M_{pre} = 0.71$ ,  $SD_{pre} = 0.29$ ,  $M_{post} = 0.91$ ,  $SD_{post} = 0.09$ ,  $\beta = 1.29$ ,  $SE = 0.11$ ,  $z = 11.96$ ,  $p < 0.01$ ) and at pre-test Mandarin learners outperformed naïve participants ( $M_{np} = 0.56$ ,  $SD_{np} = 0.27$ ,  $M_{mlp} = 0.93$ ,  $SD_{mlp} = 0.03$ ,  $\beta = 3.17$ ,  $SE = 0.09$ ,  $z = 36.09$ ,  $p < 0.001$ ). There is also a *test-session* by *participant-condition* interaction ( $\beta = -1.76$ ,  $SE = 0.21$ ,  $z = -8.54$ ,  $p < 0.01$ ). Post-hoc analysis suggested that the effect of session was only significant for naïve participants ( $\beta = 2.50$ ,  $SE = 0.18$ ,  $z = 13.31$ ,  $p < 0.01$ ), not for Mandarin learners ( $\beta = 0.61$ ,  $SE = 0.11$ ,  $z = 5.39$ ,  $p < 0.001$ ), however it will be observed that their performance is near ceiling. There was still a difference between these two groups at post-test ( $\beta = 1.41$ ,  $SE = 0.18$ ,  $z = 7.80$ ,  $p < 0.001$ ), although the difference was bigger at pre-test.

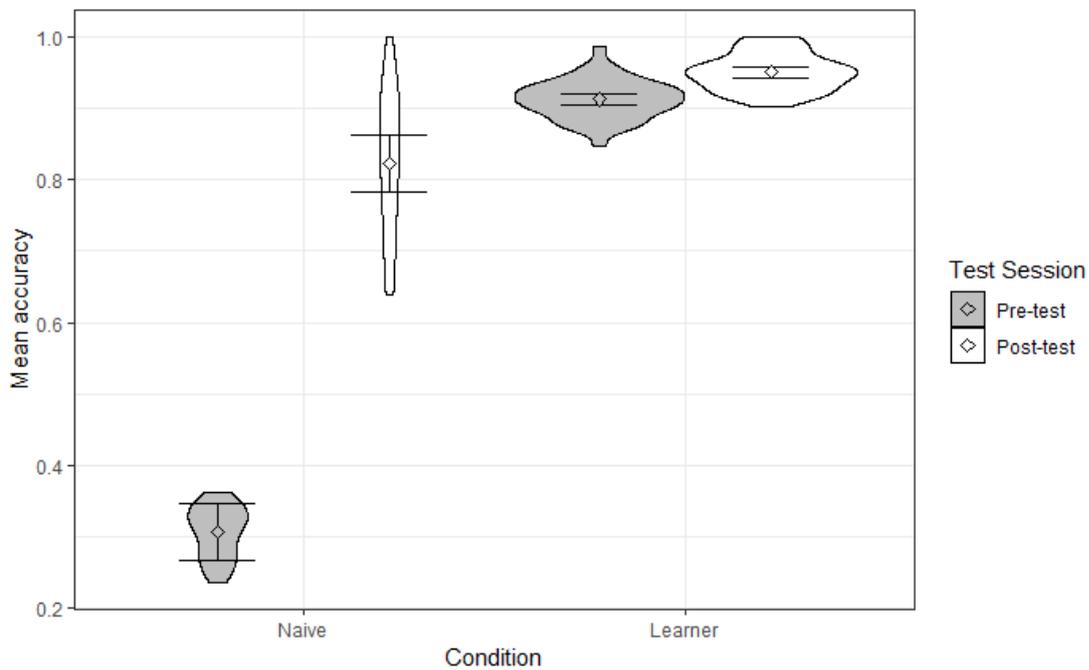


Figure 30 Mean tone accuracy of Pinyin reading for naïve participants and Mandarin learners across pre- and post-test.

#### 4.4.2.2.2 Individual differences analyses

##### 4.4.2.2.2.1 Hypotheses that the ID measure predicts performance at pre-test and that this differs for the participant groups

Relevant statistics are shown in Table 18. It can be seen that none of the *ID measures* predicted participants' performance on Pinyin Reading, tone accuracy at pre-test, with evidence for the null in each case. Age also did not predict performance, with evidence for the null. In addition, I did not find any *ID measure* x *participant-condition* interaction, however here, where I were able to compute Bayes factors, the evidence was ambiguous.

Table 18 Regression and Bayesian analysis for Tone accuracy, tone accuracy, with the effect of *ID measure* and *ID measure* x *participant-condition*, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results.

	Effect of individual aptitude at pre-test						Effect of individual aptitude by condition at pre-test (positive $\beta$ indicates larger effect in the MLP group)					
	$\beta$	SE	p	H1	Bayes	Robustness Region	$\beta$	SE	p	H1	Bayes (two tailed)	Robustness Region
Age*	0.03	0.04	0.498	0.216	0.105	[-0.10, -∞]	-0.01	0.06	0.829			

PCPT	-0.02	0.06	0.716	0.359	0.118	[0.15, ∞]	-0.04	0.10	0.681			
Working memory-composite score	-0.01	0.02	0.562	0.183	0.074	[0.05, ∞]	0.02	0.04	0.664			
Digit Span - Forward	-0.02	0.03	0.501	0.339	0.055	[0.20, ∞]	0.03	0.06	0.622			
Digit Span - Backward	-0.01	0.03	0.715	0.298	0.075	[0.25, ∞]	0.05	0.05	0.369			
Arithmetic	-0.03	0.03	0.176	0.241	0.046	[0.15, ∞]	-0.02	0.04	0.621			
Letter - Number Sequencing	-0.01	0.02	0.652	0.265	0.126	[0.10, ∞]	0.01	0.04	0.796	0.01	0.971	[0,0.20]
Attention-composite score	0.01	0.02	0.765	0.136	0.211	[0.10, ∞]	0.02	0.03	0.658	0.007	0.984	[0,0.10]
Elevator Counting with Distraction	0.01	0.03	0.860	2.968	0.117	[1.06, ∞]	0.04	0.05	0.382	0.005	0.999	[0,0.20]
Elevator Counting with Reversal	0.02	0.05	0.656	0.341	0.204	[0.25, ∞]	0.03	0.07	0.721	0.021	0.966	[0,0.20]
Visual Elevator	-0.07	0.12	0.522	-0.744	0.278	[-0.66, ∞]	-0.12	0.18	0.519	-0.075	0.953	[0,-0.61]
Telephone Search	0.06	0.15	0.69	-1.15	0.098	[-0.35, ∞]	-0.00	0.24	0.996			
Telephone Search while Counting*	-0.01	0.09	0.875	-0.316	0.324	[-0.31, ∞]	-0.0001	0.14	0.999	-0.015	0.995	[0,-0.40]
Musical ability-composite score	-0.01	0.04	0.715	0.334	0.095	[0.10, ∞]	0.03	0.08	0.671			
Beat Perception	-0.01	0.04	0.804	0.339	0.102	[0.10, ∞]	0.00	0.08	0.962			
Melody Memory	-0.01	0.04	0.783	0.44	0.08	[0.35, ∞]	0.05	0.08	0.527			

\*These ID measures were analysed with a larger number of steps (500) to make sure H1 was covered in the Robustness regions calculated.

#### 4.4.2.2.2.2 Hypotheses that ID measure predicts pre to post-test improvement, and this differs for the participant groups

Table 19 Regression and Bayesian analysis for Tone accuracy, tone accuracy, with the effect of ID measure x test-session and ID measure x test-session x participant-condition, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results.

	Effect of individual aptitude by test-session (positive $\beta$ indicates larger effect in post-test)						Effect of individual aptitude by test-session by condition (positive $\beta$ indicates larger effect in the MLP group)					
	$\beta$	SE	p	H1	Bayes	Robustness Region	$\beta$	SE	p	H1	Bayes	Robustness Region
Age	-0.07	0.08	0.405	-0.256	0.177	[-0.152, - $\infty$ ]	0.07	0.13	0.610			
PCPT	0.14	0.12	0.265	0.396	0.901	[0,1.16]	0.26	0.22	0.252	0.138	1.019	[0,1.21]
Working memory-composite score *	<b>0.10</b>	<b>0.04</b>	<b>0.014</b>	<b>0.206</b>	<b>6.387</b>	<b>[0.05,0.45]</b>	-0.17	<b>0.07</b>	<b>0.025</b>	<b>0.099</b>	<b>2.973</b>	<b>[0, 1.00]</b>
Working memory-composite score <i>NP group only</i>	<b>0.21</b>	<b>0.06</b>	<b>&lt;0.001</b>	<b>0.206</b>	<b>227.18</b>	<b>[0.05,&gt;5]</b>						
Working memory-composite score <i>MLP group only</i>	0.04	0.05	0.386	0.213	0.51	[0, 0.30]						
Digit Span - Forward	0.06	0.06	0.375	0.383	0.391	[0,0.40]	-0.17	0.13	0.196	0.057	1.047	[0,0.15]
Digit Span - Backward	<b>0.14</b>	<b>0.05</b>	<b>0.007</b>	<b>0.335</b>	<b>11.074</b>	<b>[0.10,0.35]</b>	-0.47	<b>0.10</b>	<b>1.67e-06</b>	<b>0.141</b>	<b>1429.404</b>	<b>[0.05&lt;5]</b>
Digit Span - Backward <i>NP group only</i>	<b>0.45</b>	<b>0.07</b>	<b>&lt;0.001</b>	<b>0.335</b>	<b>78719761</b>	<b>[0.05,&gt;5]</b>						
Digit Span - Backward <i>MLP group only</i>	-0.01	0.069	0.838	0.454	0.131	[0.20, $\infty$ ]						
Arithmetic	0.07	0.05	0.160	0.268	0.905	[0,0.76]	-0.06	0.09	0.525	0.074	0.845	[0,0.30]
Letter - Number Sequencing	0.09	0.05	0.046	0.293	2.084	[0.1.92]	-0.01	0.09	0.934	0.093	0.693	[0,0.20]
Attention	-0.003	0.05	0.950	0.140	0.294	[0.13, $\infty$ ]	0.01	0.08	0.920			
Elevator Counting with Distraction	-0.04	0.06	0.525	0.335	0.123	[0.15, $\infty$ ]	-0.29	0.10	0.006			
Elevator Counting with Reversal	0.02	0.10	0.879	0.372	0.292	[0.35, $\infty$ ]	0.11	0.16	0.496	0.015	0.998	[0,0.56]
Visual Elevator	-0.05	0.25	0.836	-0.804	0.349	[0,-0.81]	-0.12	0.40	0.764	-0.052	0.992	[0,-1.16]
Telephone Search	-0.17	0.33	0.605	-1.289	0.392	[0,-1.52]	-0.26	0.54	0.636	-0.171	0.964	[0,-1.72]
Telephone Search	-0.04	0.19	0.845	-0.347	0.560	[0,-0.61]	-0.30	0.30	0.311	-0.037	1.000	[0,-1.41]

while Counting												
Musical ability	0.07	0.09	0.418	0.377	0.483	[0,0.56]	-0.10	0.18	0.552	0.071	0.951	[0,0.56]
Beat Perception *	0.04	0.09	0.639	0.380	0.340	[0,0.38]	0.18	0.18	0.324	0.041	0.999	[0,0.81]
Melody Memory	0.06	0.09	0.486	0.498	0.337	[0.51,∞]	-0.41	0.18	0.020	0.062	1.264	[0,0.15]

\*These ID measures were analysed with a larger number of steps (500) to make sure H1 was covered in the Robustness regions calculated.

Relevant statistics are summarised in Table 19. There is substantial evidence that age did not predict participants' improvement from pre- to post- test. For the *Pitch Contour Perception Test*, there was ambiguous evidence that it predicted the improvement of participants, and whether there was a difference between the NP and MLP group was also ambiguous. For working memory measures, the composite score and individual *Digit Span Backward* score predicted the improvements of participants, such that the higher they score with these measures, the more they improved from pre to post test in Pinyin reading. However, there was also evidence (strong for *Digit Span Backward*, near substantial for the composite score) that this differed by *participant-condition*. Further analysis revealed that there was only evidence that these working memory measures predicted the improvement for the NP group; for the MLP group, the evidence was ambiguous for the composite score (Figure 31) though there was strong evidence for the null regarding *Digit Span Backward* (Figure 32). For attention measures, there was substantial evidence that the composite score did not predict pre to post-improvements. This was also found for two of the individual measures, *Elevator Counting with Distraction* and *Elevator Counting with Reversal*. For all other measures the evidence was ambiguous. In addition, the evidence for an interaction with participant-condition was ambiguous in every case. For musical ability measures, the composite score and each of the separate measures only demonstrated ambiguous evidence regarding whether they predicted participants' improvements, or whether there was an interaction with participant-condition.

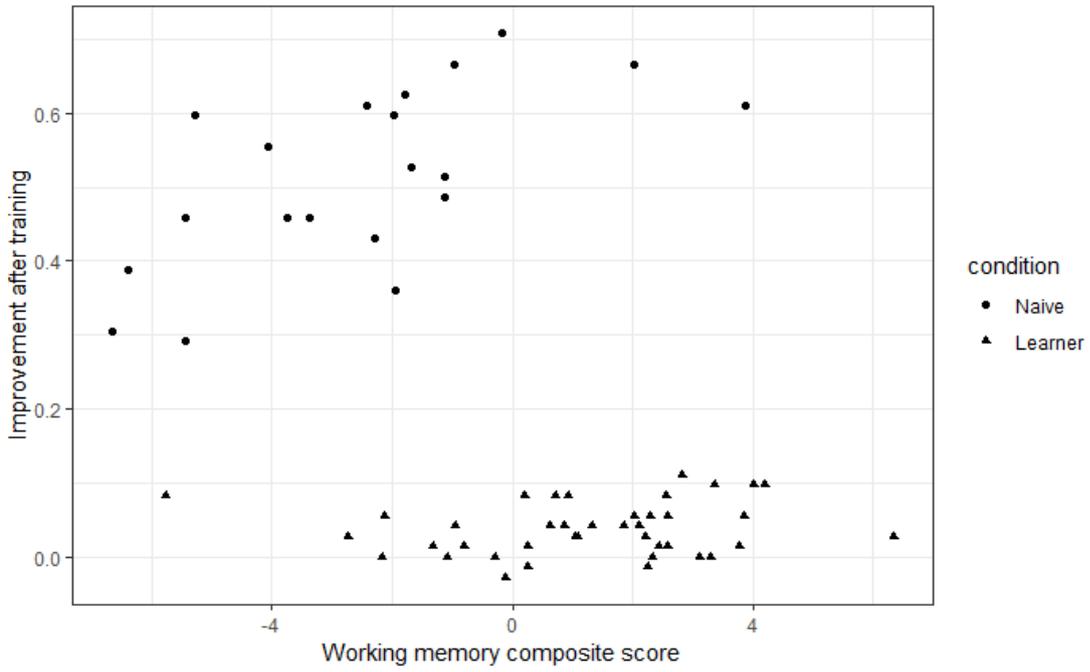


Figure 31 Scatter plot depicting the relationship between participants performance in the Working memory composite score and their improvements from pre- to post- test in the Pinyin reading tone accuracy task

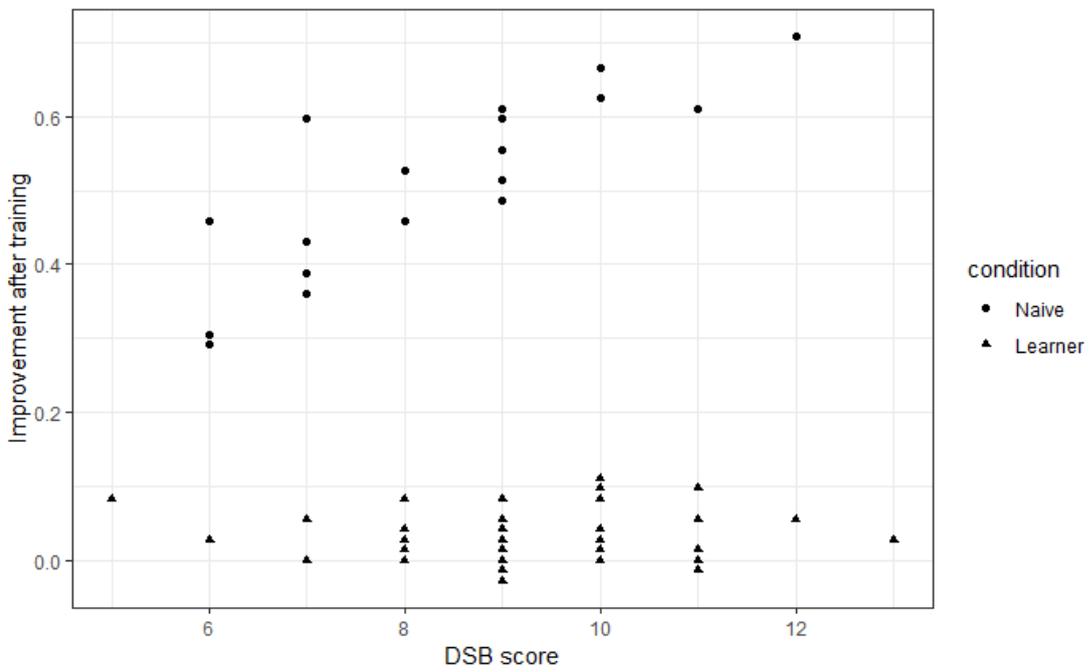


Figure 32 Scatter plot depicting the relationship between participants performance in the Digit span backwards task and their improvements from pre- to post- test in the Pinyin reading tone accuracy task

#### 4.4.2.2.3 Pinyin accuracy

##### 4.4.2.2.3.1 Analysis of performance (without *ID measures*)

The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were *test-session* (pre-test, post-test) and *participant-condition* (naïve, learner). The mean accuracy is displayed in Figure 33.

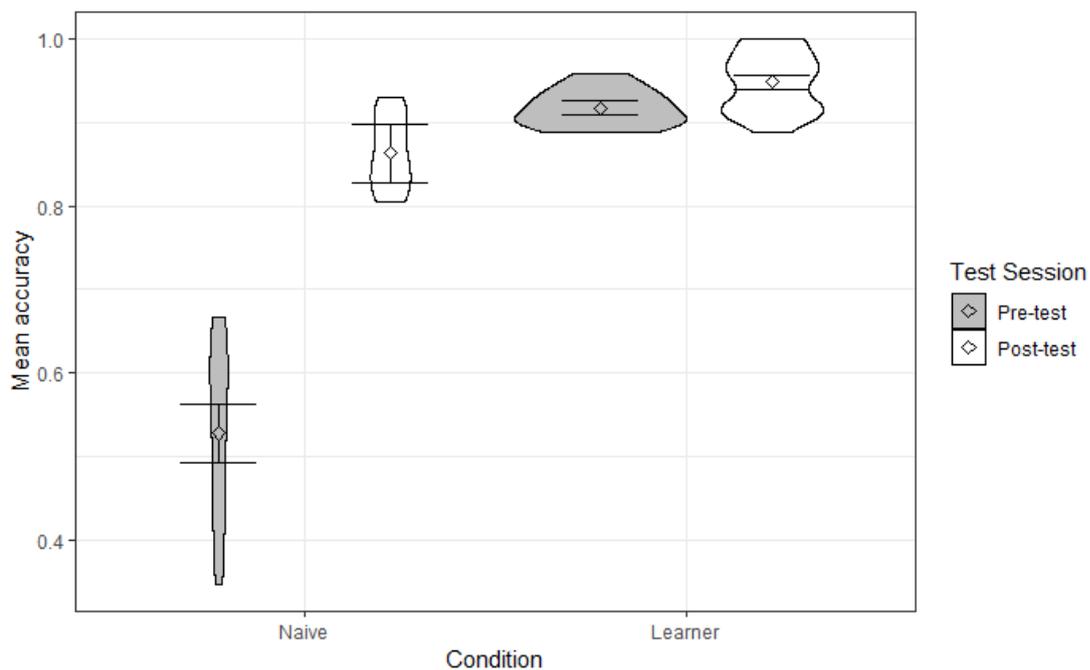


Figure 33 Mean pinyin accuracy of Pinyin reading for naïve participants and Mandarin learners across pre- and post-test.

Overall, participants performed better after training ( $M_{pre} = 0.79$ ,  $SD_{pre} = 0.19$ ,  $M_{post} = 0.92$ ,  $SD_{post} = 0.06$ ,  $\beta = 0.94$ ,  $SE = 0.09$ ,  $z = 10.24$ ,  $p < 0.001$ ) and Mandarin learners outperformed naïve participants ( $M_{np} = 0.70$ ,  $SD_{np} = 0.19$ ,  $M_{mlp} = 0.93$ ,  $SD_{mlp} = 0.03$ ,  $\beta = 2.31$ ,  $SE = 0.11$ ,  $z = 21.93$ ,  $p < 0.01$ ). There is a test-session by participant-condition interaction ( $\beta = -1.22$ ,  $SE = 0.17$ ,  $z = -7.26$ ,  $p < 0.001$ ). Post-hoc analysis suggested that the increase in performance of naïve participants after training was significant ( $\beta = 1.73$ ,  $SE = 0.12$ ,  $z = 14.98$ ,  $p < 0.01$ ) but for the Mandarin learners it was not ( $\beta = 0.63$ ,  $SE = 0.14$ ,  $z = 4.38$ ,  $p < 0.001$ ).

There was still a difference between these two groups at post-test ( $\beta = 1.09$ ,  $SE = 0.14$ ,  $z = 7.60$ ,  $p < 0.01$ ), although it was smaller than it at pre-test.

#### 4.4.2.2.3.2 Individual differences analyses.

##### 4.4.2.2.3.2.1 Hypotheses that the ID measure predicts performance at pre-test and that this differs for the participant groups

Relevant statistics are summarised in Table 20. For age, there was substantial evidence that it did not predict participants' performance at pre-test. The same was true for *Pitch Contour Perception Test*. For working memory measures, the evidence that the composite score predicted pre-test Pinyin production was ambiguous. However, there was mixed evidence for each measure separately: there was evidence for a positive relationship for *Digit Span Backward*, ambiguous evidence for *Digit Span Forward*, and evidence for the null for all other measures. However, the effect for *Digit Span Backward* was modulated by an interaction with *participant-condition* and there was a similar interaction for the composite score. Further analysis revealed that, for both the working memory composite score (Figure 34) and *Digit Span Backward* (Figure 35), there was only evidence that the ID measure predicted the performance of the NP group, such that the better they performed in these *ID measures*, the better they performed in pinyin production in Pinyin reading. For the MLP group, the evidence was ambiguous. In all other cases where a Bayes Factor was computed for the *ID measure x participant-condition* interaction, it showed the evidence was ambiguous. For attention measures, there was evidence for the null for the composite score and for *Elevator Counting with Reversal*, *Visual Elevator* and *Telephone Search*. For *Elevator Counting with Distraction* and *Telephone Search while Counting*, the evidence was ambiguous. The evidence for an interaction with *participant-condition* was ambiguous in every case. For the musical ability measures, I found clear evidence for the null for the composite score and for each of the *Beat*

*perception* and *Melody Memory* tests individually. Evidence for an ID measure x participant-condition interaction was ambiguous in every case where a Bayes Factor was computed.

Table 20 Regression and Bayesian analysis for Pinyin reading, pinyin accuracy, with the effect of ID measure and ID measure x participant-condition

Task	Effect of individual aptitude at pre-test						Effect of individual aptitude by condition at pre-test (positive $\beta$ indicates larger effect in the MLP group)					
	B	SE	p	H1	Bayes	Robustness Region	$\beta$	SE	p	H1	Bayes (2 tails)	Robustness Region
Age	-0.01	0.04	0.726	-	0.129	[-0.10,-∞]	0.01	0.07	0.945			
PCPT	-0.02	0.07	0.721	0.279	0.183	[0.15,∞]	-0.01	0.12	0.927			
Working memory composite score	0.03	0.02	0.152	0.150	0.758	[0,0.30]	<b>-0.10</b>	<b>0.04</b>	<b>0.009</b>	<b>0.032</b>	<b>3.061</b>	<b>[0.05,0.35]</b>
Working memory composite score <i>NP group only</i>	<b>0.1-</b>	<b>0.03</b>	<b>&lt;0.001</b>	<b>0.045</b>	<b>104.734</b>	<b>[0.06,3.28]</b>						
Working memory composite score <i>MLP group only</i>	-0.002	0.03	0.945	0.100	0.284	[0,0.05]						
Digit Span - Forward	0.04	0.03	0.221	0.273	0.436	[0,0.35]	-0.05	0.07	0.438	0.040	0.927	[0,0.25]
Digit Span - Backward*	<b>0.09</b>	<b>0.03</b>	<b>0.004</b>	<b>0.237</b>	<b>15.719</b>	<b>[0.05,0.35]</b>	<b>-0.17</b>	<b>0.05</b>	<b>0.002</b>	<b>0.087</b>	<b>16.647</b>	<b>[0.07,0.61]</b>
Digit Span - Backward <i>NP group only</i>	<b>0.20</b>	<b>0.35</b>	<b>&lt;0.001</b>	<b>0.057</b>	<b>,&gt;9999</b>	<b>[0.05,&gt;5]</b>						
Digit Span - Backward <i>MLP group only</i>	0.03	0.04	0.454	0.199	0.420	[0,0.25]						
Arithmetic	0.001	0.03	0.960	0.192	0.152	[0.10,∞]	-0.08	0.05	0.117	0.001	1.000	[0,0.45]
Letter - Number Sequencing	-5.84e-05	0.02	0.998	0.213	0.114	[0.10,∞]	-0.10	0.05	0.035			
Attention Composite Score	0.002	0.03	0.913	0.107	0.261	[0.10, ∞]	-0.002	0.04	0.958	0.003	0.997	[0,0.10]
Elevator Counting with Distraction	0.04	0.03	0.286	0.237	0.417	[0,0.25]	-0.05	0.05	0.350	0.036	0.952	[0,0.20]
Elevator Counting with Reversal	-0.03	0.05	0.597	0.268	0.137	[0.15,∞]	0.003	0.09	0.964			
Visual Elevator	-0.001	0.14	0.992	0.591	0.231	[-0.40,-∞]	0.03	0.22	0.873	-0.001	1.000	[0,-0.60]
Telephone Search	-0.02	0.18	0.904	-0.904	0.210	[-0.56,-∞]	0.06	0.29	0.821	-0.021	0.997	[0,-0.81]

Telephone Search while Counting	-0.02	0.11	0.859	-0.287	0.761	[0,-0.35]	-0.09	0.17	0.602	-0.02	0.995	[0,-0.51]
Musical ability Composite score	0.001	0.05	0.984	0.270	0.172	[0.16,∞]	-0.09	0.09	0.302	0.001	1.041	[0,0.40]
Beat Perception	-0.02	0.05	0.644	0.265	0.131	[0.10,∞]	0.03	0.09	0.777			
Melody Memory	0.02	0.05	0.561	0.365	0.218	[0.25,∞]	-0.18	0.09	0.043	0.028	1.149	[0,0.05]

\*These ID measures were analysed with a larger number of steps (500) to make sure H1 was covered in the Robustness regions calculated.

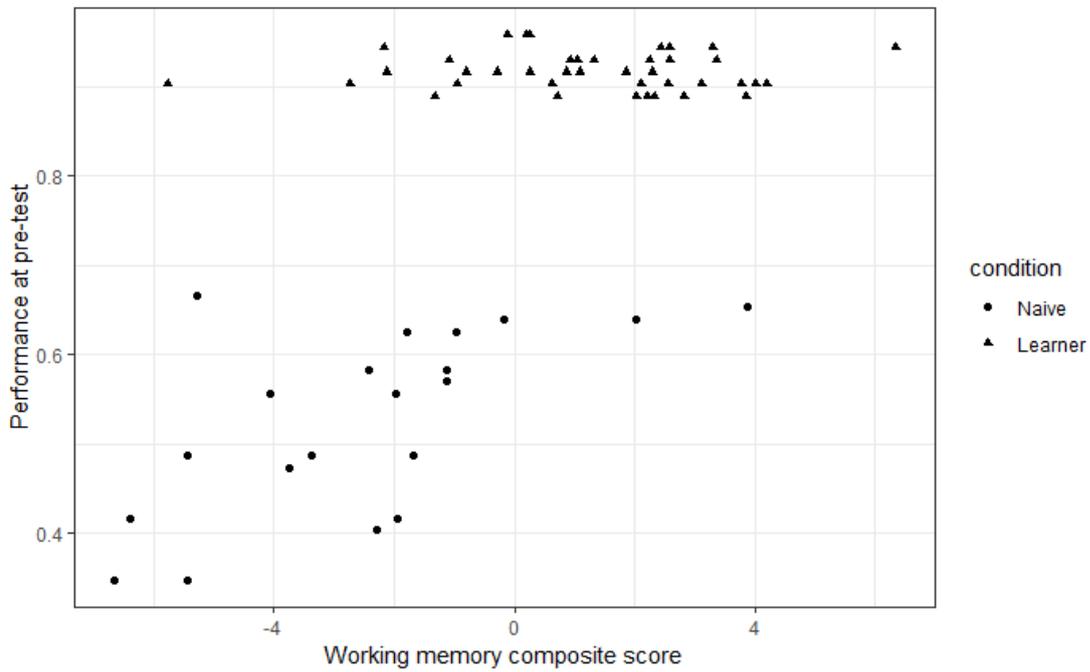


Figure 34 Scatter plot for the Pinyin reading pinyin accuracy with working memory composite score as x-axis and pre-test performance as y-axis.

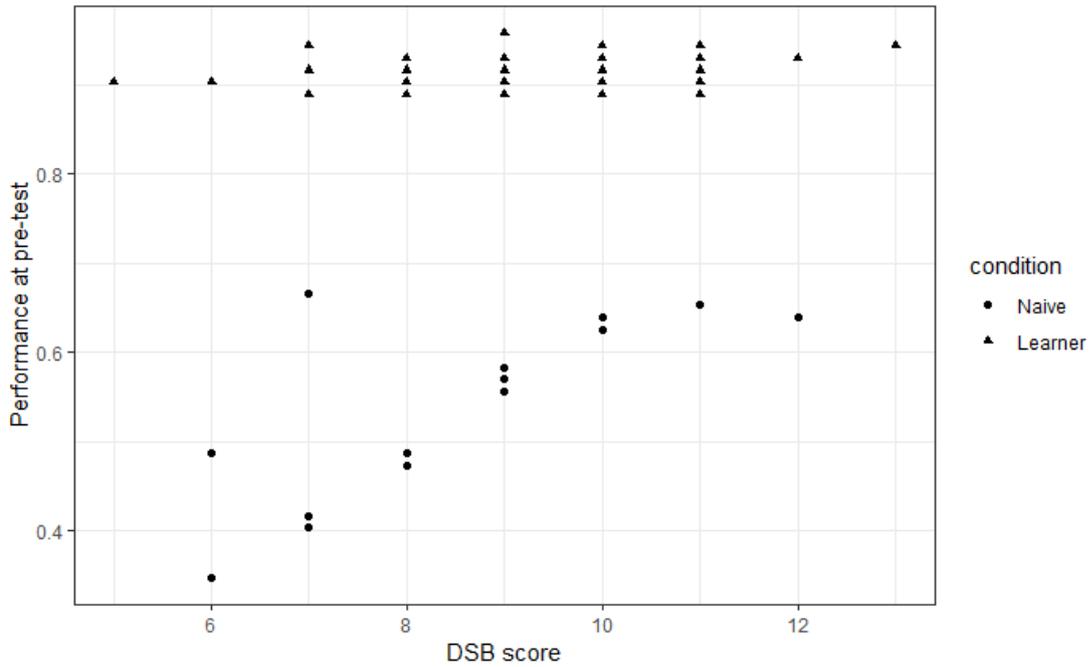


Figure 35 Scatter plot for the Pinyin reading pinyin accuracy with working memory composite score as x-axis and pre-test performance as y-axis.

4.4.2.2.3.2.2 Hypotheses that ID measure predicts pre to post-test improvement, and this differs for the participant groups

Relevant statistics are shown in Table 21. For age, there was evidence for the null that it did not predict the participants' improvement, but whether there was an interaction with participant-condition was ambiguous. For the *Pitch Contour Perception Test*, the evidence that it predicted the improvement from pre- to post- training is ambiguous, as is the evidence that this differed for different participant groups. For working memory measures, the evidence for the null was substantial for the composite score and for each measure separately. However, the evidence for an interaction was ambiguous in every case. For attention measures, there was evidence for the null for the composite score. There was also evidence for the null for the *Elevator Counting with Distraction* and *Telephone Search while Counting* measures, but for all other separate measures (*Elevator Counting with Reversal*, *Visual Elevator & Telephone Search while Counting*), the evidence is ambiguous. Similarly, the evidence was ambiguous

for an interaction with participant-condition for every measure. For musical ability measures, I found evidence for the null of the composite score and for *Beat perception* but it was ambiguous for *Melody Memory*. All evidence for *ID measure x test-session x participant-condition* remained ambiguous.

Table 21 Regression and Bayesian analysis for Pinyin reading, pinyin accuracy, with the effect of *ID measure x test-session* and *ID measure x test-session x participant-condition*, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results.

Task	Effect of individual aptitude by test-session (positive $\beta$ indicates larger effect in post-test)						Effect of individual aptitude by test-session by condition (positive $\beta$ indicates larger effect in the MLP group)					
	$\beta$	SE	p	H1	Bayes	Robustness Region	$\beta$	SE	p	H1	Bayes (2tails)	Robustness Region
Age	0.02	0.07	0.778	0.251	0.324	[0.25,∞]	0.01	0.11	0.894	0.019	0.985	[0,0.30]
PCPT	0.04	0.11	0.672	0.332	0.436	[0,0.40]	0.05	0.19	0.796	0.045	0.973	[0,0.51]
Working memory-composite score	-0.04	0.04	0.332	0.156	0.132	[0.10,∞]	-0.001	0.07	0.941			
Digit Span - Forward	-0.05	0.05	0.397	0.300	0.099	[0.10, ∞]	-0.16	0.11	0.139			
Digit Span - Backward	-0.11	0.06	0.039	0.236	0.076	[0.20,∞]	0.05	0.10	0.623			
Arithmetic	0.01	0.05	0.781	0.224	0.257	[0.20,∞]	0.01	0.08	0.856	0.013	0.987	[0,0.20]
Letter - Number Sequencing	0.02	0.04	0.697	0.249	0.234	[0.20, ∞]	0.06	0.08	0.415	0.016	0.992	[0,0.25]
Attention-composite score	0.02	0.04	0.546	0.253	0.271	[0.25, ∞]	0.002	0.06	0.971	0.024	0.936	[0,0.15]
Elevator Counting with Distraction	-0.06	0.06	0.331	0.262	0.113	[0.10, ∞]	0.01	0.09	0.914			
Elevator Counting with Reversal	0.13	0.08	0.141	0.320	1.297	[0,1.36]	0.12	0.13	0.364	0.125	0.886	[0,0.56]
Visual Elevator	-0.34	0.21	0.113	-0.703	1.721	[0,-4.19]	-0.33	0.33	0.321	-0.338	0.901	[0,-1.57]
Telephone Search	0.09	0.28	0.747	-1.067	0.204	[-0.66,-∞]	0.36	0.45	0.427			
Telephone Search while Counting	-0.09	0.16	0.585	-0.287	0.761	[0,-0.76]	0.03	0.25	0.914	-0.09	0.942	[0,-0.71]
Musical ability-composite score	-0.02	0.08	0.748	0.315	0.189	[0.20,∞]	0.02	0.14	0.888			
Beat Perception	0.01	0.08	0.943	0.318	0.250	[0.25,∞]	0.09	0.15	0.546	0.005	1.000	[0,0.50]
Melody Memory	-0.05	0.08	0.529	0.403	0.126	[0.15,∞]	-0.08	1.51	0.614			

#### 4.4.2.3 Four Interval Oddity Task

##### 4.4.2.3.1 Analysis of performance (without ID measures)

The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were *test-session* and (pre-test, post-test) *participant condition* (naive, learner). The mean accuracy is displayed in Figure 36.

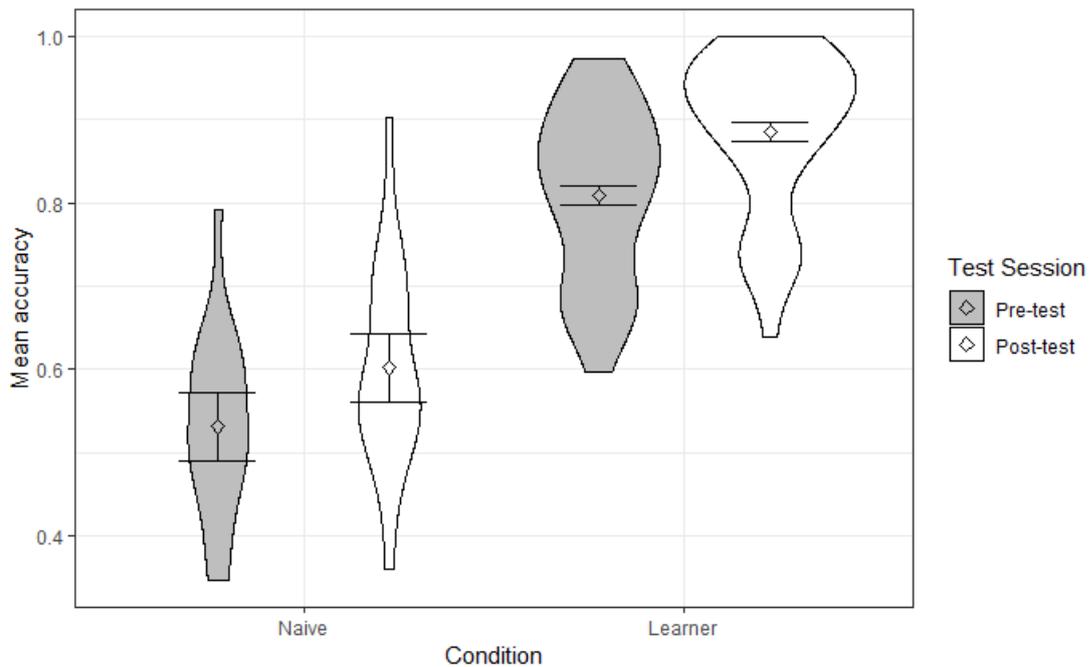


Figure 36 Mean proportion of correct of Four Interval Oddity task for naïve and learner groups in pre/post sessions,

Overall, participants performed better after training ( $M_{pre} = 0.72$ ,  $SD_{pre} = 0.17$ ,  $M_{post} = 0.79$ ,  $SD_{post} = 0.17$ ,  $\beta = 0.64$ ,  $SE = 0.08$ ,  $z = 7.92$ ,  $p < 0.001$ ) and Mandarin learners outperformed naïve participants ( $M_{np} = 0.57$ ,  $SD_{np} = 0.12$ ,  $M_{mlp} = 0.85$ ,  $SD_{mlp} = 0.11$ ,  $\beta = 1.43$ ,  $SE = 0.17$ ,  $z = 8.30$ ,  $p < 0.001$ ). There is an interaction between *test-session* and *participant-condition* ( $\beta = 0.48$ ,  $SE = 0.15$ ,  $z = 3.15$ ,  $p < 0.01$ ). Post-hoc analysis suggested that both groups increased after training but the improvement was more notable for Mandarin learners ( $\beta = 0.83$ ,  $SE = 0.11$ ,  $z = 7.57$ ,  $p < 0.01$ ) than for naïve participants ( $\beta = 0.31$ ,  $SE =$

0.12,  $z = 2.58$ ,  $p = 0.01$ ). There was still a difference between these two groups at post-test ( $\beta = 1.87$ ,  $SE = 0.23$ ,  $z = 7.99$ ,  $p < 0.01$ ) but the difference was smaller compared with pre-test.

#### 4.4.2.3.2 Individual difference analysis

##### 4.4.2.3.2.1 Hypotheses that the *ID measure* predicts performance at pre-test and that this differs for the participant groups

The results of the Bayesian analysis are summarised in Table 22. For age there was substantial evidence that it did not predict performance at pre-test. There was strong evidence that the *Pitch Contour Perception Test* predicted the performance at pre-test, although there was also strong evidence suggesting an interaction with *participant-condition* (Figure 37). Further analyses suggested that the *Pitch Contour Perception Test* was only predictive for the MLP group and there was evidence for the null for the NP group. For working memory measures, strong evidence was found that both the composite score and the *Letter Number Sequencing* predicted pre-test performance. The effect of *Letter Number Sequencing* was also modulated by an interaction with *participant-condition* (Figure 38). After breaking down, there was a strong evidence for an effect of *Letter Number Sequencing* only in the MLP group, with the evidence for the null in the NP group. For all other working memory measures, I only found ambiguous evidence for both the effect of *ID measure* and the interaction with *participant-condition*. For attention, I found strong evidence for the composite score (Figure 39) and for each of the separate measures: *Elevator Counting with Reversal* (Figure 40), *Visual Elevator* (Figure 41) & *Telephone Search while Counting* (Figure 42), except for *Elevator Counting with Distraction* where the evidence was ambiguous. However, in each case, there was strong evidence for an interaction with *participant-condition*. Further analysis suggested that there was only evidence that these tasks predicted the performance of the MLP group at pre-test, with the evidence for the null in the NP group. For *Elevator Counting with Distraction* the evidence for the interaction was also ambiguous. For musical ability measures, I found strong

evidence that the composite score predicted pre-test performance but the separate measures patterned differently: I found strong evidence for *Beat perception* but evidence for the null for *Melody Memory*. The evidence for an interaction with participant-condition was ambiguous.

Table 22 Regression and Bayesian analysis for Four Interval Oddity task, with the effect of *ID measure* and *ID measure x participant-condition*, with green cells representing evidence for *H1*, red cells representing evidence for the Null and yellow cells representing ambiguous results.

Task	Effect of individual aptitude at pre-test						Effect of individual aptitude by condition at pre-test (positive $\beta$ indicates larger effect in the MLP group)					
	B	SE	p	H1	Bayes	Robustness Region	$\beta$	SE	p	H1	Bayes (2 tails)	Robustness Region
Age	-0.08	0.06	0.196	-	0.127	[0,-0.25]	0.14	0.10	0.176			
PCPT	<b>0.46</b>	<b>0.07</b>	<b>&lt;0.001</b>	<b>0.236</b>	<b>&gt;9999</b>	[0.05,>5]	<b>0.67</b>	<b>0.13</b>	<b>&lt;0.001</b>	<b>0.465</b>	<b>&gt;9999</b>	[0.10,>5]
PCPT <i>NP group only</i>	0.09	0.02	0.831	0.688	0.158	[0.35, $\infty$ ]						
PCPT <i>MLP group only</i>	<b>0.69</b>	<b>0.09</b>	<b>&lt;0.001</b>	<b>0.259</b>	<b>&gt;9999</b>	[0.05,>5]						
Working memory-composite score	<b>0.11</b>	<b>0.03</b>	<b>0.953</b>	<b>0.161</b>	<b>62.278</b>	[0.05,1.21]	0.12	0.07	0.073	0.107	1.695	[1.01, $\infty$ ]
Digit Span - Forward	0.09	0.05	0.064	0.314	1.645	[1.67, $\infty$ ]	-0.02	0.11	0.888	0.093	0.775	[0,0.30]
Digit Span - Backward	0.05	0.05	0.297	0.271	0.539	[0,0.40]	0.08	0.11	0.420	0.054	0.952	[0,0.40]
Arithmetic	0.03	0.04	0.413	0.223	0.398	[0,0.25]	-0.06	0.08	0.410	0.034	0.967	[0,0.30]
Letter - Number Sequencing	<b>0.14</b>	<b>0.03</b>	<b>&lt;0.001</b>	<b>0.23</b>	<b>&gt;9999</b>	[0.05,>5]	<b>0.28</b>	<b>0.06</b>	<b>&lt;0.001</b>	<b>0.145</b>	<b>5828.98</b>	[0.05,>5]
Letter - Number Sequencing <i>NP group only *</i>	-0.05	0.041	0.268	0.241	0.081	[0.20, $\infty$ ]						
Letter - Number Sequencing <i>MLP group only</i>	<b>0.24</b>	<b>0.04</b>	<b>&lt;0.001</b>	<b>0.28</b>	<b>&gt;9999</b>	[0.05,>5]						

Attention-composite score	0.15	0.03	<0.001	0.113	>9999	[0.05,>5]	0.18	0.05	<0.001	0.148	177.695	[0.05,>5]
Attention-composite score <i>NP group only</i>	0.03	0.02	0.062	0.208	0.885	[0,0.61]						
Attention-composite score <i>MLP group only</i>	0.21	0.05	<0.001	0.133	6037.513	[0.05,>5]						
Elevator Counting with Distraction	0.06	0.05	0.266	0.271	0.565	[0,0.45]	0.04	0.08	0.661	0.055	0.863	[0,0.25]
Elevator Counting with Reversal	0.34	0.06	<0.001	0.298	>9999	[0.05,>5]	0.46	0.10	<0.001	0.337	6037.246	[0.10,>5]
Elevator Counting with Reversal <i>NP group only</i>	0.03	0.041	0.422	0.496	0.177	[0.30,0.86]						
Elevator Counting with Reversal <i>MLP group only</i>	0.50	0.09	<0.001	0.359	>9999	[0.05,>5]						
Visual Elevator	-0.66	0.17	<0.001	-0.641	757.038	[-0.10,>5]	-0.82	0.27	0.003	-0.664	16.87	[-0.40,>-2.12]
Visual Elevator <i>NP group only</i>	-0.12	0.097	0.214	-0.945	0.39	[0,-1.06]						
Visual Elevator <i>MLP group only</i>	-0.95	0.27	<0.001	-0.761	147.828	[-0.05,>5]						
Telephone Search	-0.98	0.21	<0.001	-1.017	>9999	[-0.10,>5]	-1.03	0.35	0.003	-0.979	15.438	[-0.51,-2.32]
Telephone Search <i>NP group only</i>	-0.29	0.15	0.051	-1.334	1.42	[0,>5]						
Telephone Search <i>MLP group only</i>	-1.33	0.34	<0.001	-1.223	739.515	[-0.05,>5]						
Telephone Search while Counting	-0.46	0.13	0.0006	-0.279	99.604	[-0.10,>5]	-0.59	0.21	0.005	-0.458	11.353	[-0.40,-0.86]
Telephone Search while Counting <i>NP group only *</i>	-0.06	0.04	0.150	-0.66	0.338	[0, -0.67]						
Telephone Search while Counting	-0.66	0.22	<0.01	-0.336	22.236	[-0.20,-3.33]						

<i>MLP group only</i>												
Musical ability-composite score	0.22	0.07	<0.001	0.303	156.402	[0.05,>5]	0.05	0.13	0.728	0.222	0.543	[0,0.40]
Beat Perception	0.31	0.06	<0.001	0.302	>9999	[0.05,>5]	0.23	0.12	0.053	0.308	1.848	[0,2.32]
Melody Memory	-0.04	0.07	0.588	0.430	0.116	[0.15,∞]	-0.26	0.15	0.097			

\*These ID measures were analysed with a larger number of steps (500) to make sure H1 was covered in the Robustness regions calculated.

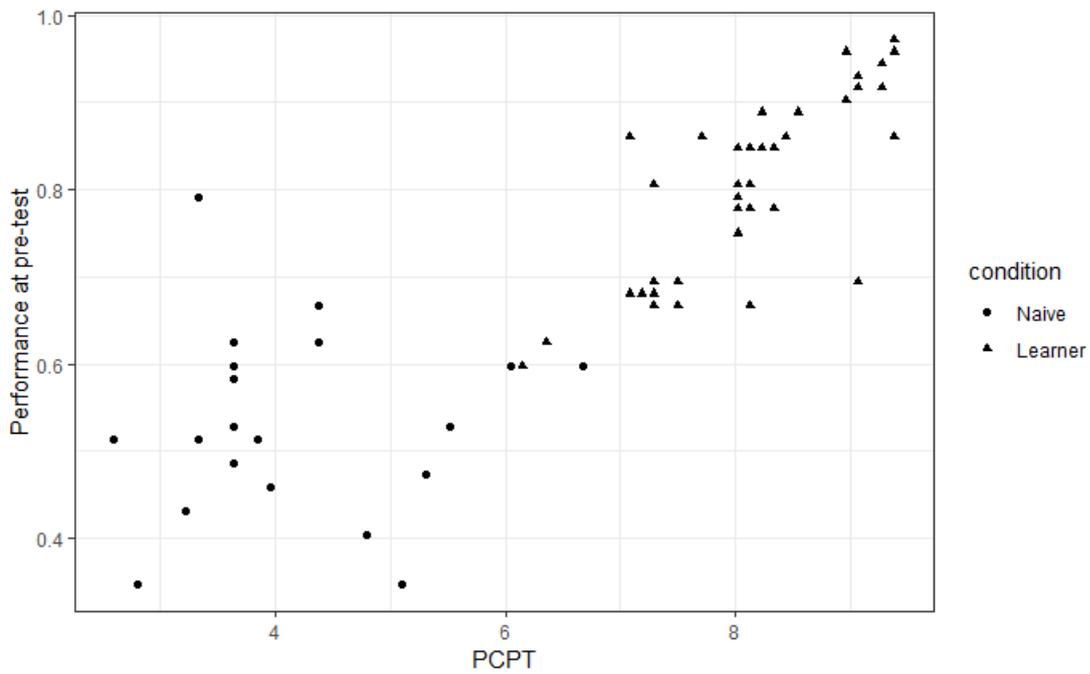


Figure 37 Scatter plot for the Four Interval Oddity task with PCPT score as x-axis and pre-test performance as y-axis.

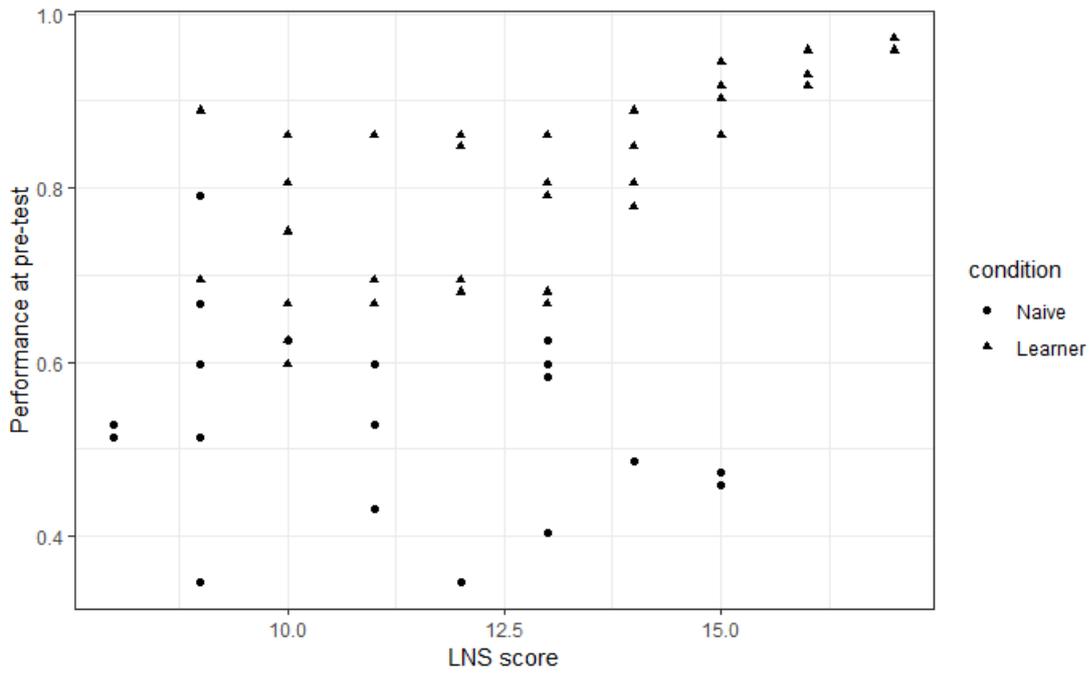


Figure 38 Scatter plot for the Four Interval Oddity task with LNS score as x-axis and pre-test performance as y-axis.

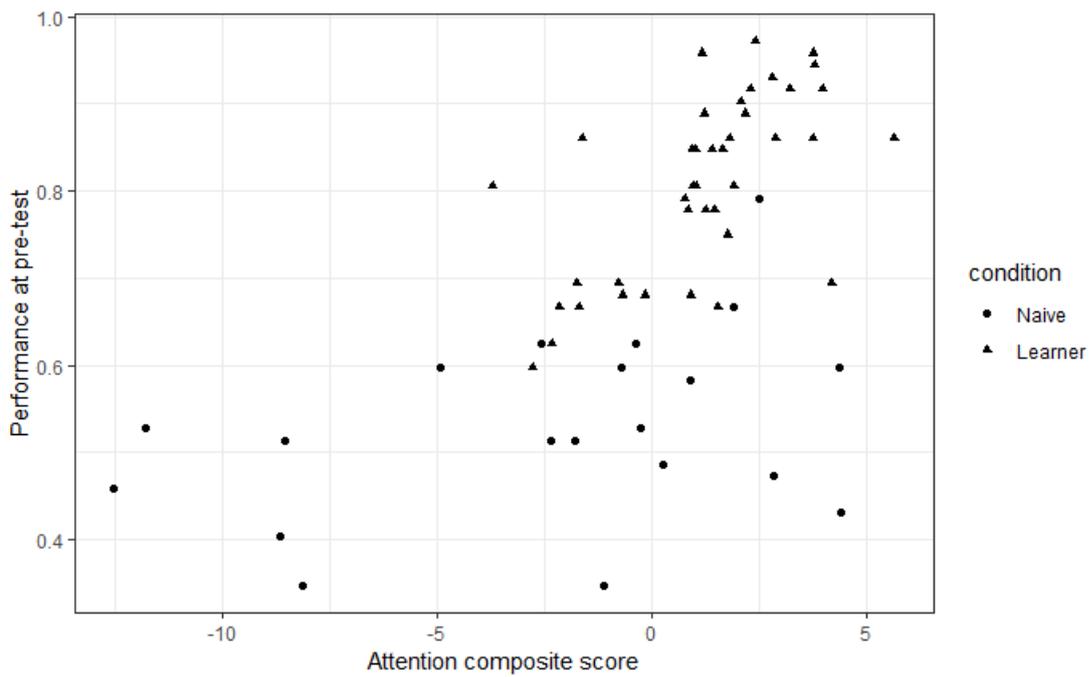


Figure 39 Scatter plot for the Four Interval Oddity task with Attention composite score as x-axis and pre-test performance as y-axis.

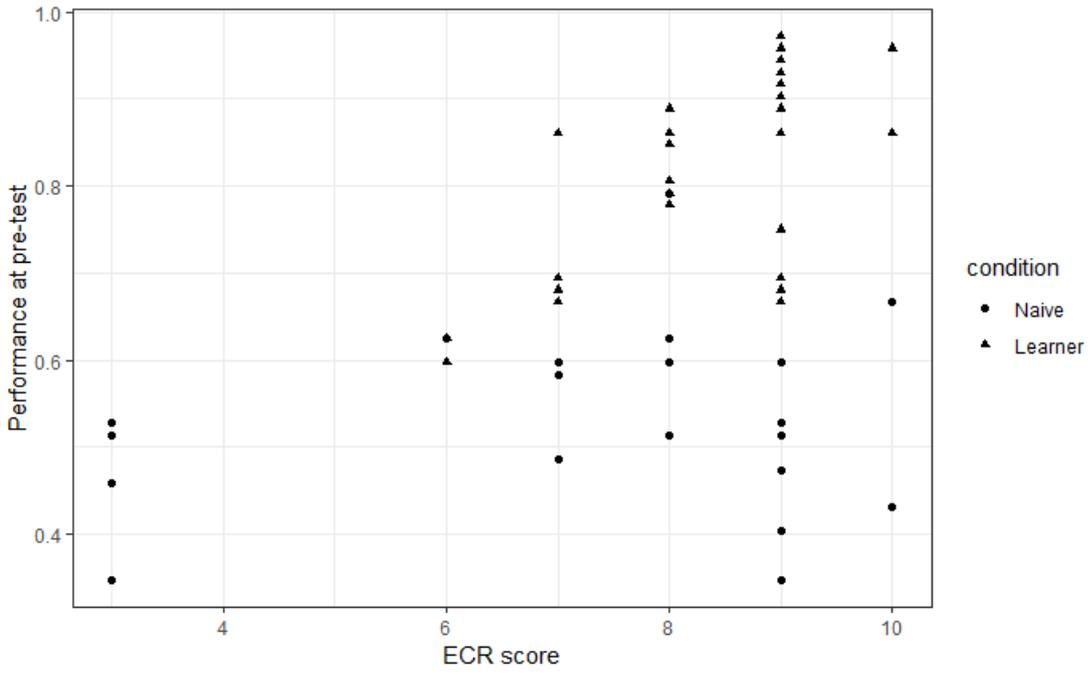


Figure 40 Scatter plot for the Four Interval Oddity task with ECR score as x-axis and pre-test performance as y-axis.

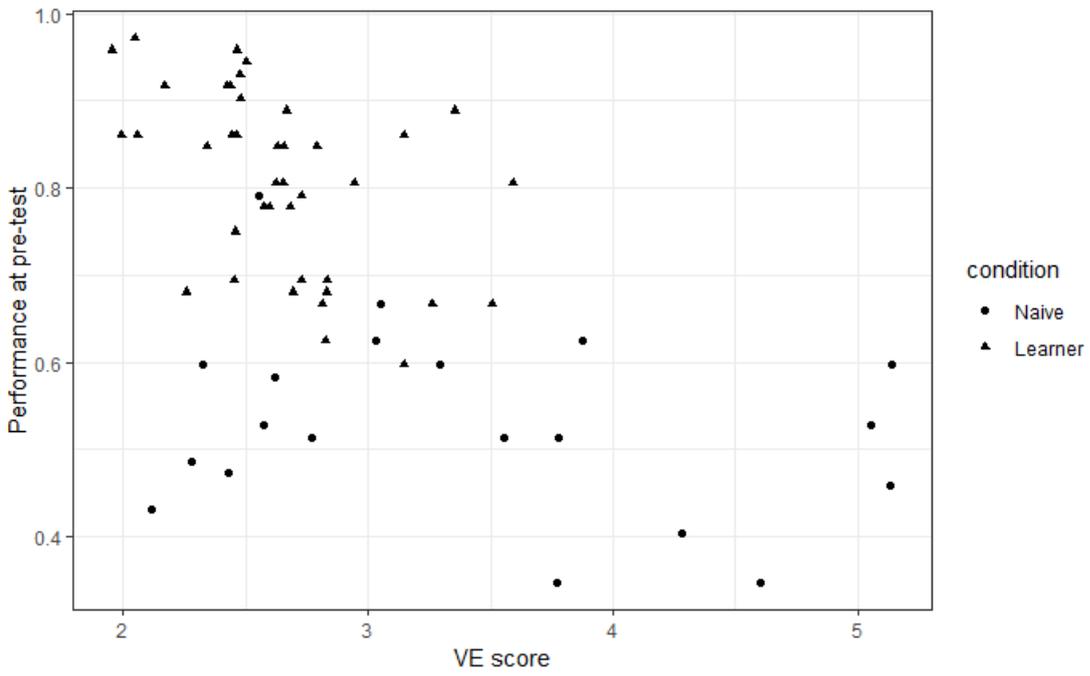


Figure 41 Scatter plot for the Four Interval Oddity task with VE score as x-axis and pre-test performance as y-axis.

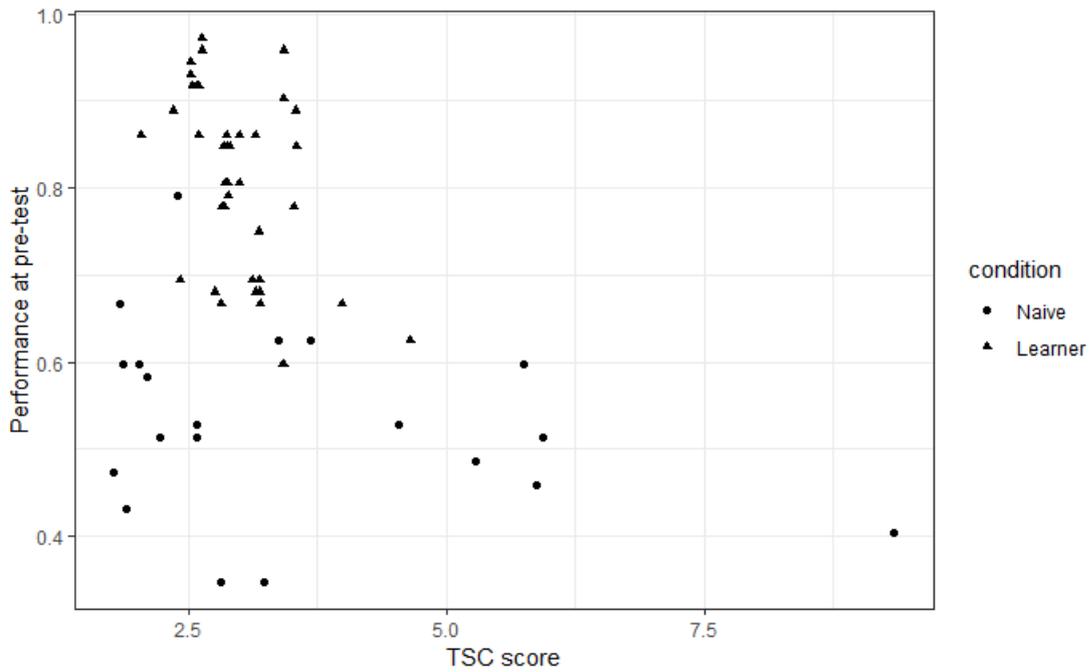


Figure 42 Scatter plot for the Four Interval Oddity task with TSC score as x-axis and pre-test performance as y-axis.

#### 4.4.2.3.2.2 Hypotheses that *ID measure* predicts pre to post-test improvement, and this differs for the participant groups

Relevant results are summarised in Table 23. The evidence regarding whether age predicts participants' improvement after training was ambiguous. For the *Pitch Contour Perception Test*, there was strong evidence that it predicted pre- to post- test improvement, but there was also an interaction with participant-condition (Figure 37). Breaking this down, there was evidence for an effect in the MLP group but the evidence was ambiguous for the NP group. For the working memory measures, there was substantial evidence that the composite score and the individual *Digit Span Forward* scores were predictive of participant's improvement. In both cases (working memory composite score (Figure 44) and *Digit Span Forward* (Figure 45)) there was substantial evidence that this was modulated by an interaction with participant-condition. Further analysis revealed that there was only evidence found for an interaction in the NP group, while the evidence remained ambiguous for the MLP group. For all other

working memory measures, the evidence was ambiguous, as was the evidence for the interaction with participant-condition. For attention measures, there was strong evidence that the composite score was predictive of participants' improvement and there was substantial evidence for the individual measures *Elevator Counting with Distraction*, *Elevator Counting with Reversal* and *Telephone Search*; for the composite score, *Elevator Counting with Reversal* and *Telephone Search* there was evidence that this was modulated by an interaction with participant-condition (this was ambiguous for *Elevator Counting with Distraction*). In each case, further analysis showed that there evidence for the null for the MLP group and evidence for the null for the NP group, (attention composite score (Figure 46), *Elevator Counting with Reversal* (Figure 47), *Telephone Search* (Figure 48)). For TSC the evidence that its prediction at pre-test was ambiguous but there was evidence for an interaction with participant condition. However, further analysis suggested that there was only evidence for the null for the NP group and the evidence that it was predictive was ambiguous for the MLP groups (Figure 49). For musical ability, the evidence that the composite score predicted improvement was ambiguous, as was the evidence that this differed across groups. A similar pattern was found *Beat perception* (evidence for both the *ID measure x test-session* and *ID measure x test-session x participant-condition* interactions was ambiguous). *Melody Memory* showed evidence that it predicted participant's improvement and there was also evidence suggesting a group difference (Figure 50). After breaking down, *Melody Memory* only predicted the improvement of the NP group, with the evidence for the MLP group remaining ambiguous.

Table 23 Regression and Bayesian analysis for Four Interval Oddity, with the effect of *ID measure x test-session* and *ID measure x test-session x participant-condition*, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results.

Task	Effect of individual aptitude by test-session (positive $\beta$ indicates larger effect in post-test)						Effect of individual aptitude by test-session by condition (positive $\beta$ indicates larger effect in the MLP group)					
	B	SE	p	H1	Bayes	Robustness Region	B	SE	p	H1	Bayes (2 tails)	Robustness Region
Age	-0.09	0.06	0.135	-0.282	0.088	[0,-0.05]	0.11	0.10	0.260			

PCPT	<b>0.33</b>	<b>0.08</b>	< <b>0.00</b> <b>1</b>	<b>0.164</b>	<b>677.835</b>	[ <b>0.05</b> ,> <b>5</b> ]	<b>0.33</b>	<b>0.14</b>	<b>0.021</b>	<b>0.334</b>	<b>3.774</b>	[ <b>0.20</b> , <b>0.56</b> ]
PCPT <i>NP group only</i>	0.11	0.11	0.323	0.426	0.662	[0,0.86]						
PCPT <i>MLP group only</i>	<b>0.43</b>	<b>0.10</b>	<b>0.052</b>	< <b>0.00</b> <b>1</b>	<b>9.026</b>	[ <b>0.05</b> , <b>0.05</b> ]						
Working memory composite score	<b>0.07</b>	<b>0.03</b>	<b>0.02</b>	<b>0.152</b>	<b>5.247</b>	[ <b>0.05</b> , <b>0.25</b> ]	<b>-0.13</b>	<b>0.05</b>	<b>0.013</b>	<b>0.068</b>	<b>4.359</b>	[ <b>0.05</b> , <b>0.35</b> ]
Working memory composite score <i>NP group only</i>	<b>0.15</b>	<b>0.04</b>	< <b>0.00</b> <b>1</b>	<b>0.192</b>	<b>1827.988</b>	[ <b>0.05</b> ,> <b>5</b> ]						
Working memory composite score <i>MLP group only</i>	0.04	0.05	0.385	0.147	0.659	[0,0.30]						
Digit Span - Forward	<b>0.11</b>	<b>0.04</b>	<b>0.005</b>	<b>0.336</b>	<b>11.237</b>	[ <b>0.05</b> , <b>0.35</b> ]	<b>-0.23</b>	<b>0.09</b>	<b>0.009</b>	<b>0.112</b>	<b>5.346</b>	[ <b>0.10</b> , <b>0.86</b> ]
Digit Span - Forward <i>NP group only</i>	<b>0.25</b>	<b>0.07</b>	< <b>0.00</b> <b>1</b>	<b>0.332</b>	<b>142.92</b>	[ <b>0.05</b> ,> <b>5</b> ]						
Digit Span - Forward <i>MLP group only</i>	0.04	0.05	0.415	0.248	0.458	[0,0.30]						
Digit Span - Backward	0.03	0.05	0.457	0.308	0.297	[0.30,∞]	<b>-0.20</b>	<b>0.09</b>	<b>0.020</b>	<b>0.034</b>	<b>1.326</b>	[0,0.30]
Arithmetic	0.05	0.04	0.154	0.251	0.731	[0,0.56]	<b>-0.11</b>	<b>0.06</b>	<b>0.070</b>	<b>0.052</b>	<b>1.515</b>	[0,0.96]
Letter – Number Sequencing *	0.07	0.04	0.066	0.226	1.648	[0,0.10]	<b>-0.02</b>	<b>0.07</b>	<b>0.789</b>	<b>0.067</b>	<b>0.72</b>	[0,0.15]
Attention composite score	<b>0.11</b>	<b>0.03</b>	< <b>0.00</b> <b>1</b>	<b>0.076</b>	<b>341.607</b>	[ <b>0.05</b> ,> <b>5</b> ]	<b>0.19</b>	<b>0.05</b>	< <b>0.00</b> <b>1</b>	<b>0.113</b>	<b>259.455</b>	[ <b>0.05</b> ,> <b>5</b> ]
Attention composite score <i>NP group only</i>	-0.01	0.02	0.637	0.174	0.1	[0.05,∞]						
Attention composite score <i>MLP group only</i>	<b>0.21</b>	<b>0.05</b>	< <b>0.00</b> <b>1</b>	<b>0.138</b>	<b>6392.815</b>	[ <b>0.05</b> ,> <b>5</b> ]						
Elevator Counting with Distraction	<b>0.11</b>	<b>0.04</b>	<b>0.011</b>	<b>0.308</b>	<b>6.734</b>	[ <b>0.20</b> , <b>0.71</b> ]	0.11	0.07	0.124	0.112	1.247	[0,0.66]
Elevator Counting with Reversal	<b>0.22</b>	<b>0.06</b>	<b>0.0004</b>	<b>0.214</b>	<b>171.027</b>	[ <b>0.05</b> ,> <b>5</b> ]	<b>0.39</b>	<b>0.10</b>	< <b>0.00</b> <b>1</b>	<b>0.222</b>	<b>194.658</b>	[ <b>0.10</b> ,> <b>5</b> ]
Elevator Counting with Reversal <i>NP group only</i> *	-0.03	0.05	0.495	0.343	0.09	[0.33,∞]						
Elevator Counting with Reversal <i>MLP group only</i>	<b>0.34</b>	<b>0.08</b>	< <b>0.00</b> <b>1</b>	<b>0.14</b>	<b>514.287</b>	[ <b>0.05</b> ,> <b>5</b> ]						
Visual Elevator	-0.24	0.18	0.180	-0.509	1.293	[0,2.27]	<b>-0.52</b>	<b>0.28</b>	<b>0.060</b>	<b>-0.235</b>	<b>1.556</b>	[0,4.65]

Telephone Search	<b>-0.82</b>	<b>0.21</b>	<b>0.0001</b>	<b>-0.126</b>	<b>6.041</b>	[-0.10,-0.25]	<b>-1.38</b>	<b>0.34</b>	<b>&lt;0.001</b>	<b>-0.318</b>	<b>5.177</b>	[-0.20,-1.41]
Telephone Search <i>NP group only</i>	-0.102	0.20	0.605	-1.253	0.107	[-0.40,-∞]						
Telephone Search <i>MLP group only</i>	<b>1.25</b>	<b>0.29</b>	<b>&lt;0.001</b>	<b>-0.799</b>	<b>3554.939</b>	[-0.15,>-5]						
Telephone Search while Counting	-0.32	0.13	0.013	-0.024	1.473	[0,-0.05]	<b>-0.50</b>	<b>0.20</b>	<b>0.012</b>	<b>-0.317</b>	<b>5.642</b>	[-0.202,-1.61]
Telephone Search while Counting <i>NP group only</i>	0.01	0.06	0.829	-0.487	0.11	[-0.15,-∞]						
Telephone Search while Counting <i>MLP group only</i>	-0.49	0.18	0.006	-0.025	1.382	[0,-0.05]						
Musical ability composite score	0.11	0.06	0.068	0.266	2.137	[0,1.87]	-0.003	0.12	0.982	0.112	0.742	[0, 0.30]
Beat Perception	0.05	0.06	0.425	0.226	0.548	[0,0.35]	0.29	0.12	0.021	0.049	1.323	[0,0.10]
Melody Memory*	<b>0.14</b>	<b>0.06</b>	<b>0.022</b>	<b>0.498</b>	<b>3.21</b>	[0.04,0.53]	<b>-0.36</b>	<b>0.12</b>	<b>0.002</b>	<b>0.138</b>	<b>9.558</b>	[0.08,0.14]
Melody Memory <i>NP group only</i>	<b>0.37</b>	<b>0.10</b>	<b>&lt;0.001</b>	<b>0.498</b>	<b>469.235</b>	[0.10,>5]						
Melody Memory <i>MLP group only</i>	-0.007	0.089	0.933	0.369	0.223	[0.25,∞]						

\*These ID measures were analysed with a larger number of steps (500) to make sure H1 was covered in the Robustness regions calculated.

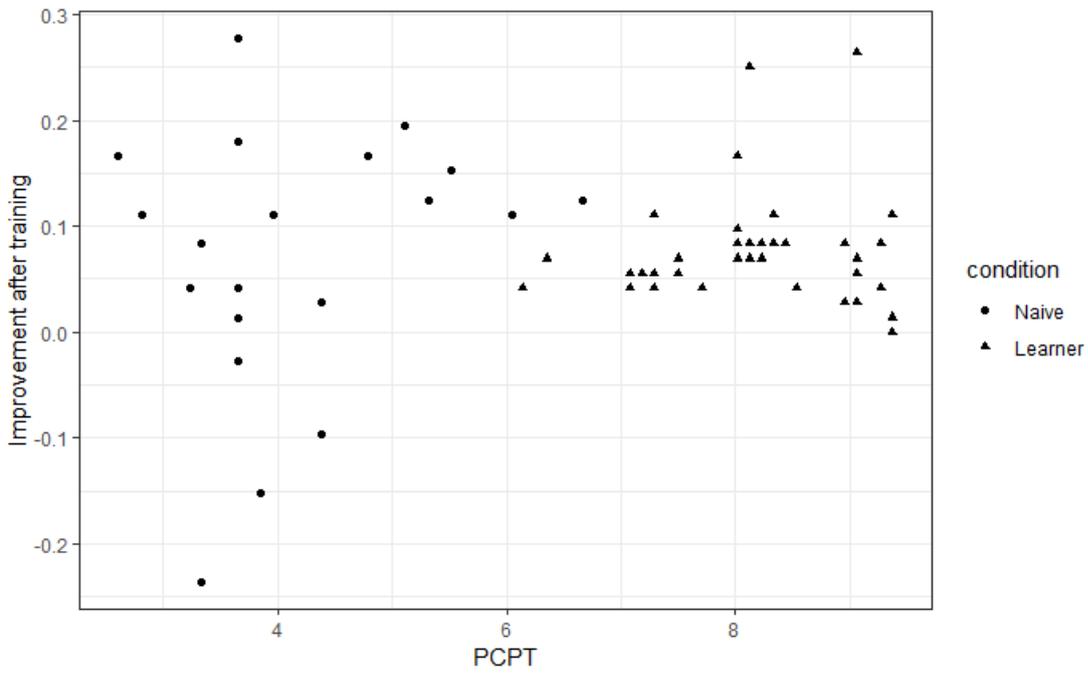


Figure 43 Scatter plot for Four-Interval Oddity measure, with PCPT as x-axis and pre/post-test difference as y-axis.

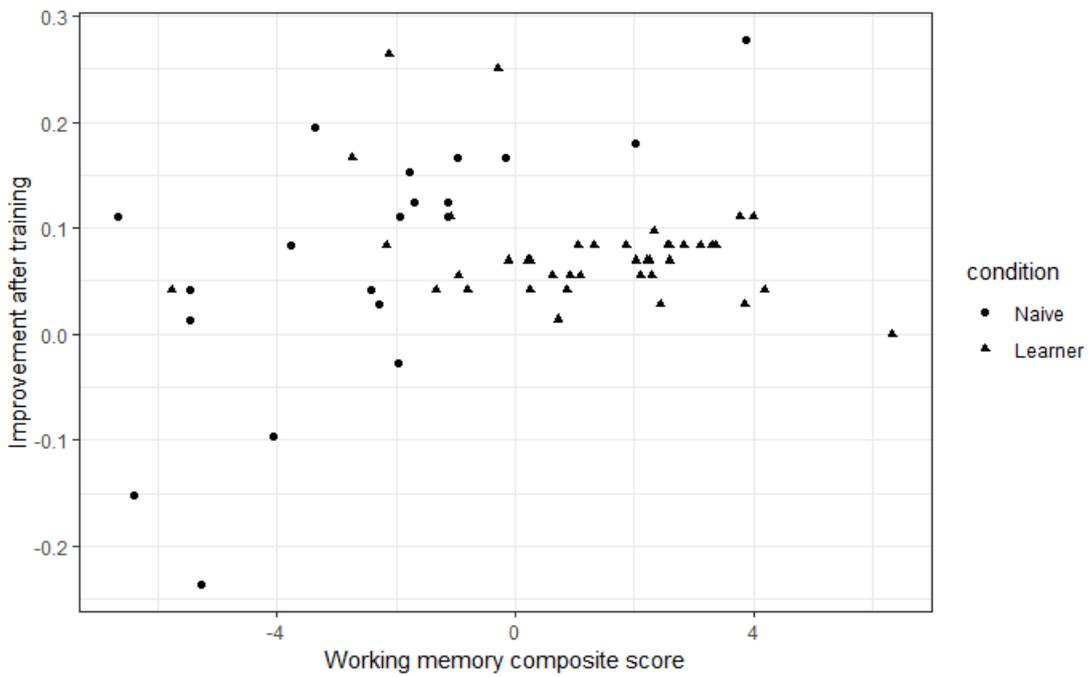


Figure 44 Scatter plot for Four-Interval Oddity measure, with Working memory composite score as x-axis and pre/post-test difference as y-axis.

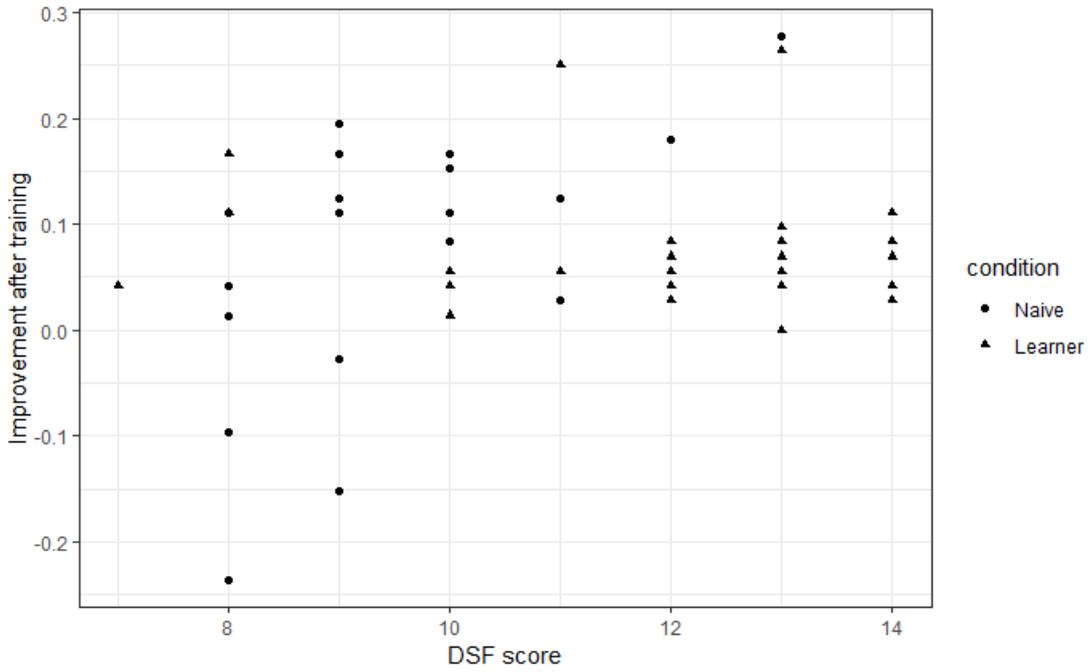


Figure 45 Scatter plot for Four-Interval Oddity measure, with Digit Span Forward score as x-axis and pre/post-test difference as y-axis.

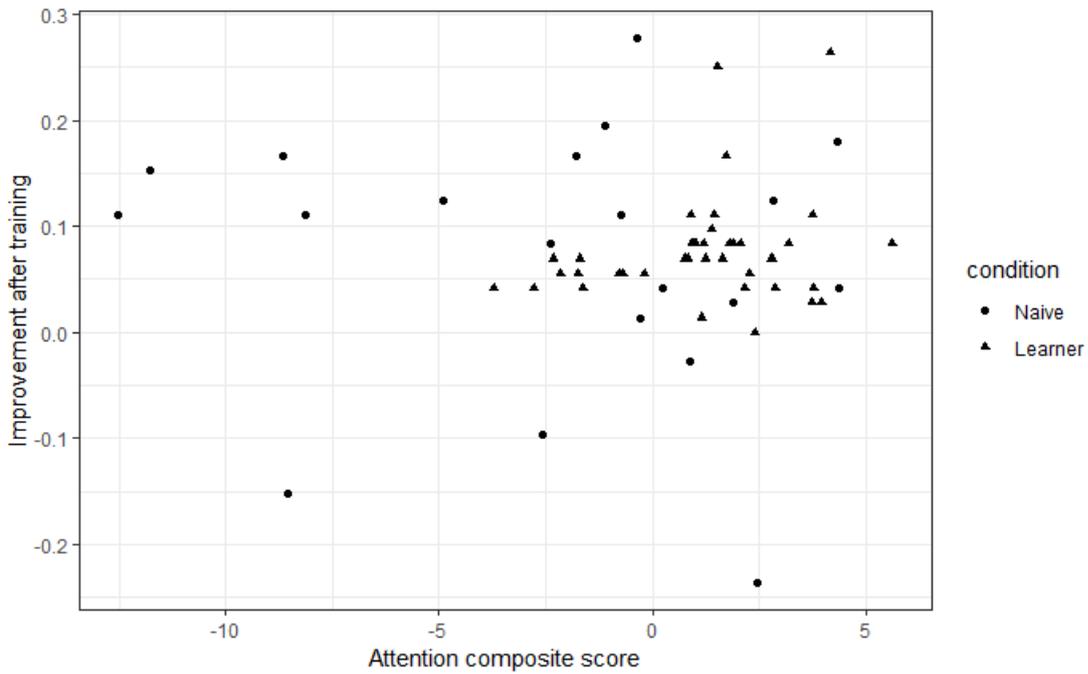


Figure 46 Scatter plot for Four-Interval Oddity measure, with Attention composite score as x-axis and pre/post-test difference as y-axis.

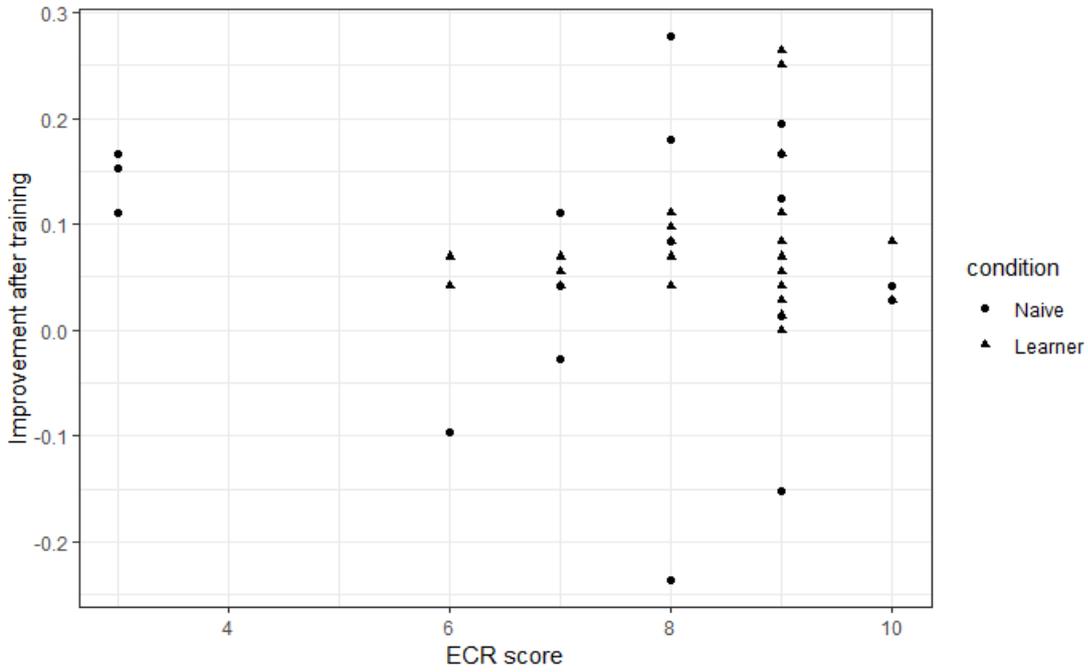


Figure 47 Scatter plot for Four-Interval Oddity measure, with Elevator Counting with Reversal score as x-axis and pre/post-test difference as y-axis.

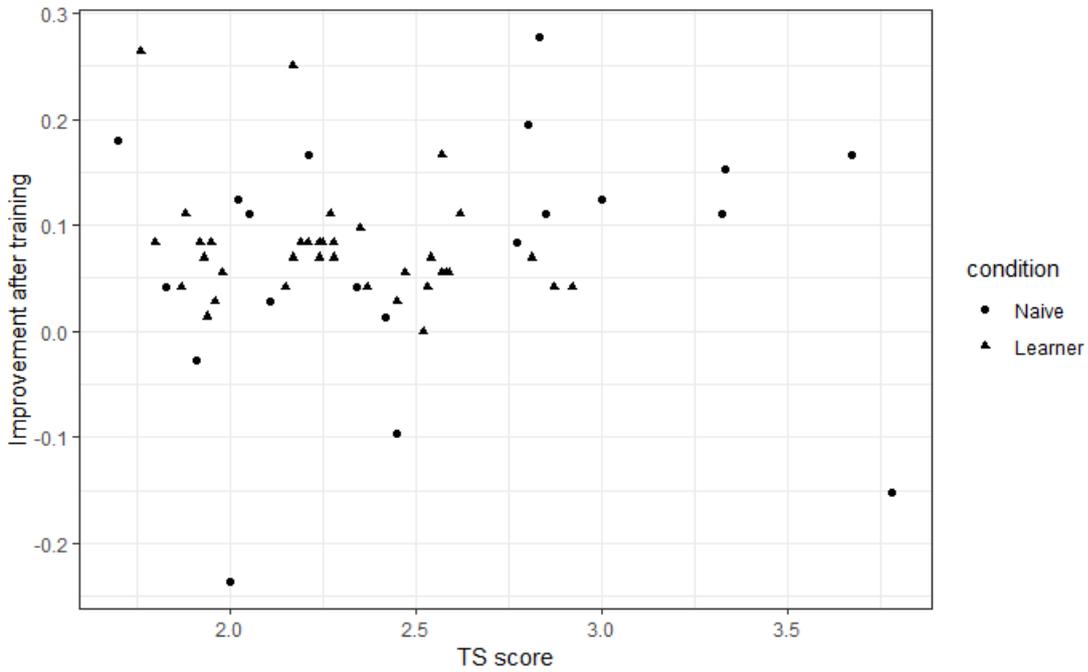


Figure 48 Scatter plot for Four-Interval Oddity measure, with Telephone Search score as x-axis and pre/post-test difference as y-axis.

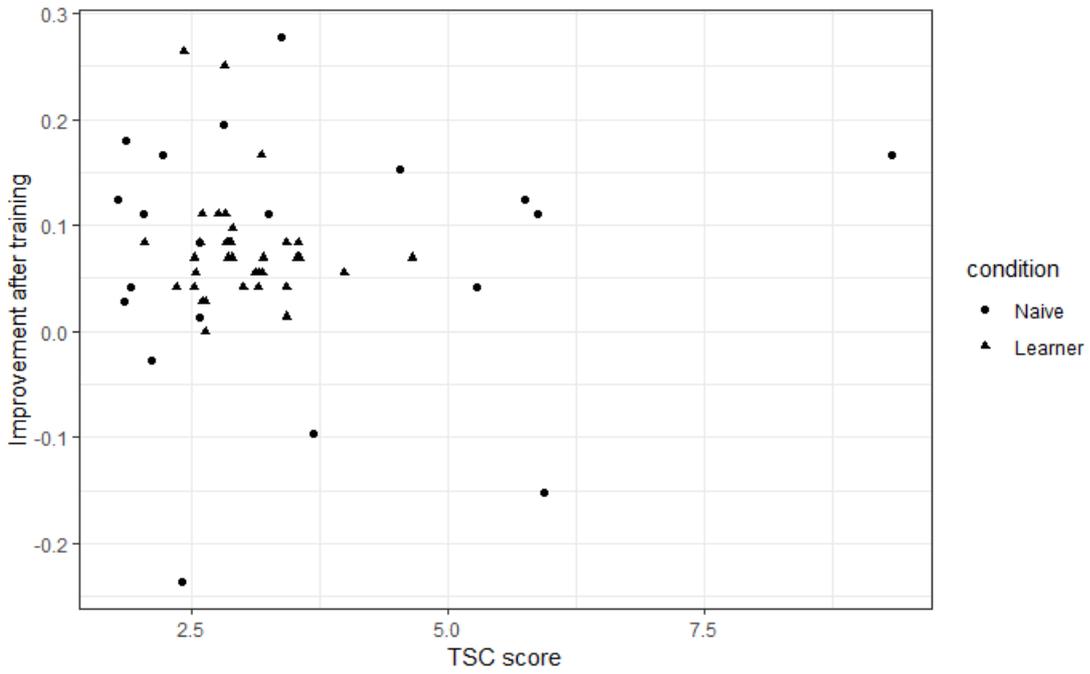


Figure 49 Scatter plot for Four-Interval Oddity measure, with Telephone Search while Counting score as x-axis and pre/post-test difference as y-axis.

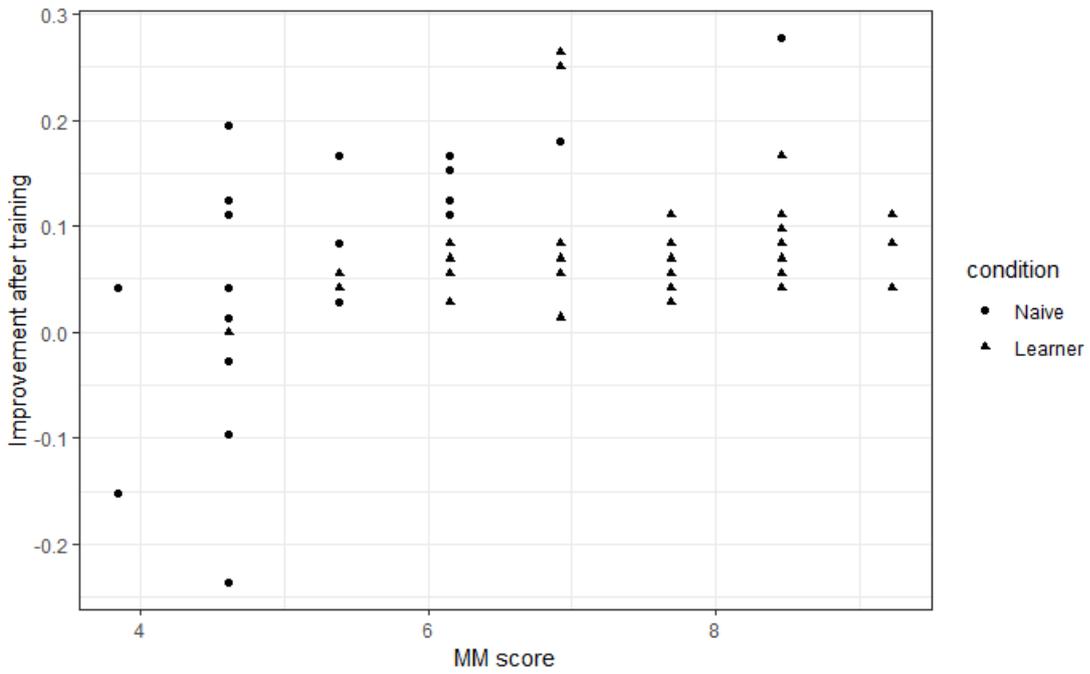


Figure 50 Scatter plot for Four-Interval Oddity measure, with Melody Memory score as x-axis and pre/post-test difference as y-axis.

#### 4.4.2.4 Pitch Contour Perception Test

##### 4.4.2.4.1 Analysis of performance (without ID measures)

The predicted variable was whether a correct response was given (1/0) on each trial. The predictors were *test-session* and (pre-test, post-test) *participant-condition* (naïve, learner). The mean accuracy is displayed in Figure 51.

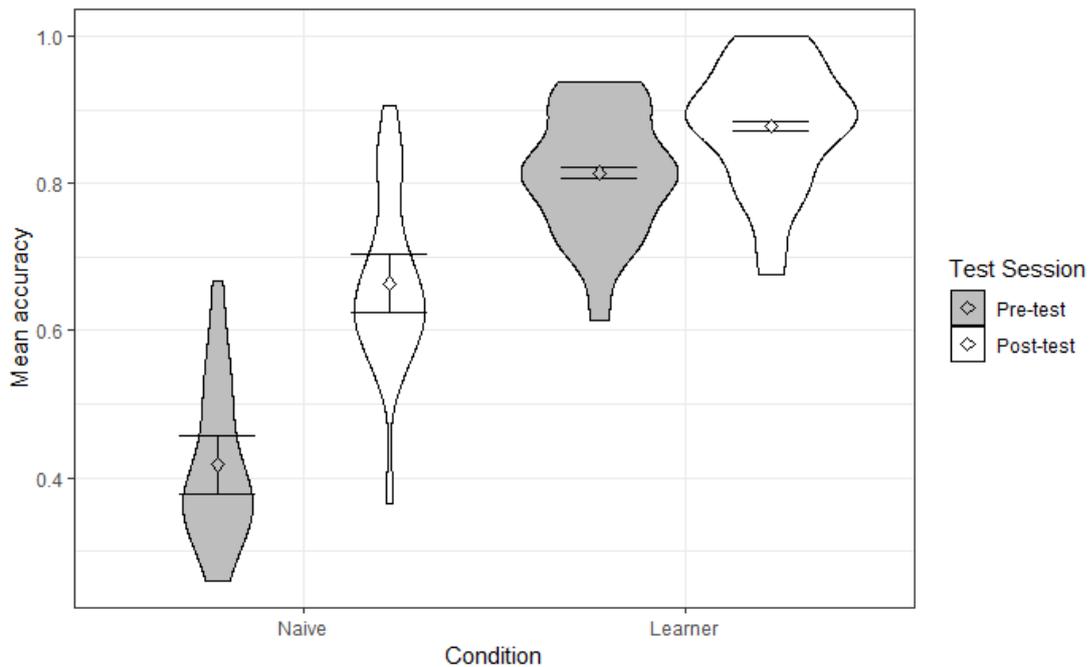


Figure 51 Mean proportion of correct of Pitch Contour Perception Test for naïve participants and Mandarin learners across pre- and post-test.

Participants performed better after training ( $M_{pre} = 0.68$ ,  $SD_{pre} = 0.21$ ,  $M_{post} = 0.81$ ,  $SD_{post} = 0.14$ ,  $\beta = 0.78$ ,  $SE = 0.07$ ,  $z = 11.43$ ,  $p < 0.001$ ) and Mandarin learners outperformed naïve participants ( $M_{np} = 0.54$ ,  $SD_{np} = 0.17$ ,  $M_{mlp} = 0.85$ ,  $SD_{mlp} = 0.09$ ,  $\beta = 1.91$ ,  $SE = 0.15$ ,  $z = 12.85$ ,  $p < 0.01$ ) at pre-test. There is an interaction between *test-session* and *participant-condition* ( $\beta = -0.43$ ,  $SE = 0.13$ ,  $z = -3.27$ ,  $p < 0.01$ ). Post-hoc analysis suggested that both groups increased after training but the improvement was more notable for naïve participants ( $\beta = 1.07$ ,  $SE = 0.12$ ,  $z = 8.93$ ,  $p < 0.01$ ) than for Mandarin learners ( $\beta = 0.65$ ,  $SE = 0.09$ ,  $z = 7.21$ ,

$p < 0.01$ ). The advantage of Mandarin learners over naïve participants still existed at post-test ( $\beta = 1.44$ ,  $SE = 0.21$ ,  $z = 6.81$ ,  $p < 0.01$ ) but was smaller compared with pre-test.

#### 4.4.2.4.2 Individual differences analyses

##### 4.4.2.4.2.1 Hypotheses that the *ID measure* predicts performance at pre-test and that this differs for the participant groups

The results of the Bayesian analysis are summarised in Table 24. There was substantial evidence that age was not predictive of participants' performance at pre-test. For working memory measures, there was strong evidence that the composite score predicted participants' performance. This was also found for *Letter Number Sequencing*. There was also a *Letter Number Sequencing x participant-condition* interaction, where there was only evidence for the effect of LNS in the MLP group, and evidence for the null in the NP group (Figure 52). For all other working memory measures, there was no effect at pre-test (evidence for the null for arithmetic, ambiguous evidence for *Digit Span Forward* and *Digit Span Backward*) or interaction with condition (ambiguous in each case). For attention measures, there was very strong evidence that the composite score was predictive of pre-test performance, and also that there was also an interaction with participant-condition (Figure 53). A similar pattern was found for *Elevator Counting with Reversal* (Figure 54), *Visual Elevator* (Figure 55), *Telephone Search* (Figure 56) & *Telephone Search while Counting* (Figure 57). Further analysis revealed that for all these measures, there was only evidence for an effect in the MLP group, and evidence for the null in the NP group. The only exception was *Elevator Counting with Distraction*, where ambiguous evidence was reported for both effect at pre-test and interaction with participant-condition. For musical ability measures, there was strong evidence that composite score was predictive at pre-test. However, *Beat perception* and *Melody Memory* showed different patterns. While there was evidence for the null for *Melody Memory* at pre-test, there was evidence that *Beat perception* was predictive and also that it interacted with

participant-condition (Figure 58). After breaking down, it was found that *Beat perception* only predicted the performance of the MLP group at pre-test, while there was evidence for the null for the NP group.

Table 24 Regression and Bayesian analysis for Pitch Contour Perception Test, with the effect of *ID measure* and *ID measure x participant-condition*, with green cells representing evidence for H1, red cells representing evidence for the Null and yellow cells representing ambiguous results.

Task	Effect of individual aptitude at pre-test						Effect of individual aptitude by condition at pre-test (positive $\beta$ indicates larger effect in the MLP group)					
	B	SE	p	H1	Bayes	Robustness Region	$\beta$	SE	p	H1	Bayes (2 tails)	Robustness Region
Age	-0.06	0.04	0.115	-0.184	0.087	[0,-0.71]						
Working memory composite score	<b>0.08</b>	<b>0.028</b>	<b>0.004</b>	<b>0.152</b>	<b>20.686</b>	[0.05,0.35]	0.07	0.06	0.222	0.081	0.95	[0,0.30]
Digit Span - Forward	0.08	0.04	0.062	0.288	1.656	[0,1.46]	0.02	0.12	0.873	0.08	0.771	[0,0.25]
Digit Span - Backward *	0.04	0.04	0.388	0.253	0.397	[0,0.30]	0.01	0.09	0.910	0.038	0.922	[0,0.25]
Arithmetic	0.01	0.04	0.839	0.206	0.2	[0.15,∞]	-0.05	0.07	0.487	0.007	0.997	[0,0.20]
Letter - Number Sequencing	<b>0.12</b>	<b>0.03</b>	<b>&lt;0.001</b>	<b>0.218</b>	<b>6774.842</b>	[0.05,>5]	<b>0.17</b>	<b>0.05</b>	<b>0.002</b>	<b>0.123</b>	<b>22.153</b>	[0.10,0.61]
Letter - Number Sequencing NP group only	0.01	0.042	0.785	0.18	0.284	[0.15,∞]						
Letter - Number Sequencing MLP group only	<b>0.18</b>	<b>0.04</b>	<b>&lt;0.001</b>	<b>0.284</b>	<b>&gt;99999</b>	[0.05,5]						
Attention composite score	<b>0.14</b>	<b>0.02</b>	<b>&lt;0.001</b>	<b>0.105</b>	<b>&gt;99999</b>	[0.05,>5]	<b>0.24</b>	<b>0.04</b>	<b>&lt;0.001</b>	<b>0.143</b>	<b>&gt;99999</b>	[0.05,>5]
Attention composite score NP group only	-0.02	-0.02	0.393	0.223	0.049	[0.15,∞]						
Attention composite score MLP group only	<b>0.22</b>	<b>0.03</b>	<b>&lt;0.001</b>	<b>0.137</b>	<b>&gt;99999</b>	[0.05,5]						
Elevator Counting with Distraction	0.06	0.04	0.121	0.248	1	[0,0.76]	0.11	0.07	0.139	0.065	1.208	[0,0.61]

Elevator Counting with Reversal	0.32	0.05	<0.001	0.277	>99999	[0.05,>5]	0.42	0.08	<0.001	0.324	>99999	[0.05,>5]
Elevator Counting with Reversal <i>NP group only</i>	0.04	0.04	0.294	0.463	0.267	[0.40,∞]						
Elevator Counting with Reversal <i>MLP group only</i>	0.46	0.07	<0.001	0.36	>99999	[0.05,5]						
Visual Elevator	-0.46	0.14	<0.001	-0.584	85.844	[-0.10,>5]	-0.97	0.23	<0.001	-0.465	649.006	[-0.15,>5]
Visual Elevator <i>NP group only</i>	0.18	0.09	0.048	-0.797	0.041	[-0.30,∞]						
Visual Elevator <i>MLP group only</i>	-0.80	0.22	<0.001	-0.773	212.507	[-0.15,>5]						
Telephone Search	-1.02	0.15	<0.001	-0.924	>99999	[-0.05,>5]	-1.76	0.26	<0.001	-1.021	>99999	[-0.10,>5]
Telephone Search <i>NP group only</i>	0.15	0.16	0.352	-1.602	0.052	[-0.86,∞]						
Telephone Search <i>MLP group only</i>	-1.60	0.21	<0.001	-1.209	>99999	[-0.05,-5]						
Telephone Search while Counting	-0.44	0.11	<0.001	-0.256	580.644	[-0.10,>5]	-0.74	0.17	<0.001	-0.435	1244.745	[-0.15,>5]
Telephone Search while Counting <i>NP group only</i>	0.04	0.04	0.264	-0.685	0.036	[-0.25,∞]						
<i>MLP</i> Telephone Search while Counting <i>group only</i>	-0.69	0.17	<0.001	-0.335	46.952	[-0.10,-5]						
Musical ability composite score	0.14	0.06	0.016	0.266	6.599	[0.10,>5]	0.275	0.12	0.021	0.135	2.971	[0.0,135]
Beat Perception	0.18	0.05	<0.001	0.273	115.794	[0.05,3.89]	0.45	0.11	<0.001	0.176	217.542	[0.10,5]
Beat Perception <i>NP group only</i>	-0.124	0.091	0.171	0.325	0.119	[0.15,∞]						
Beat Perception <i>MLP group only</i>	0.33	0.06	<0.001	0.362	>99999	[0.05,5]						
Melody Memory	-0.02	0.06	0.799	0.388	0.132	[0.15,∞]						

\*These ID measures were analysed with more intense steps to make sure H1 was covered in the Robustness regions calculated.

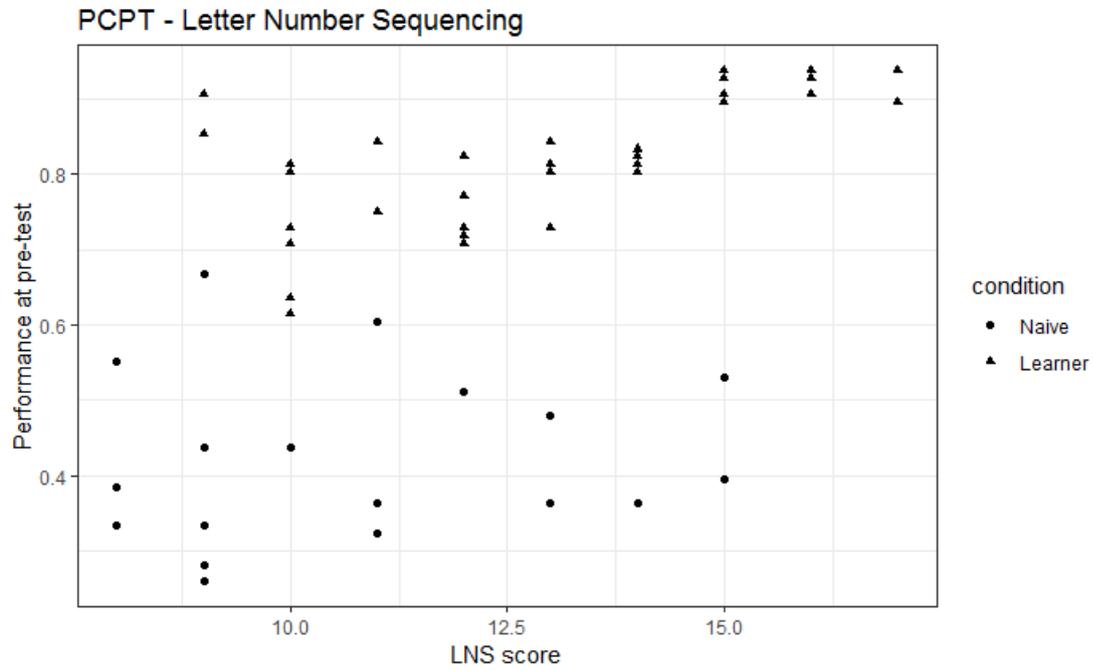


Figure 52 Scatter plot for the Pitch Contour Perception Test with LNS score as x-axis and pre-test performance as y-axis.

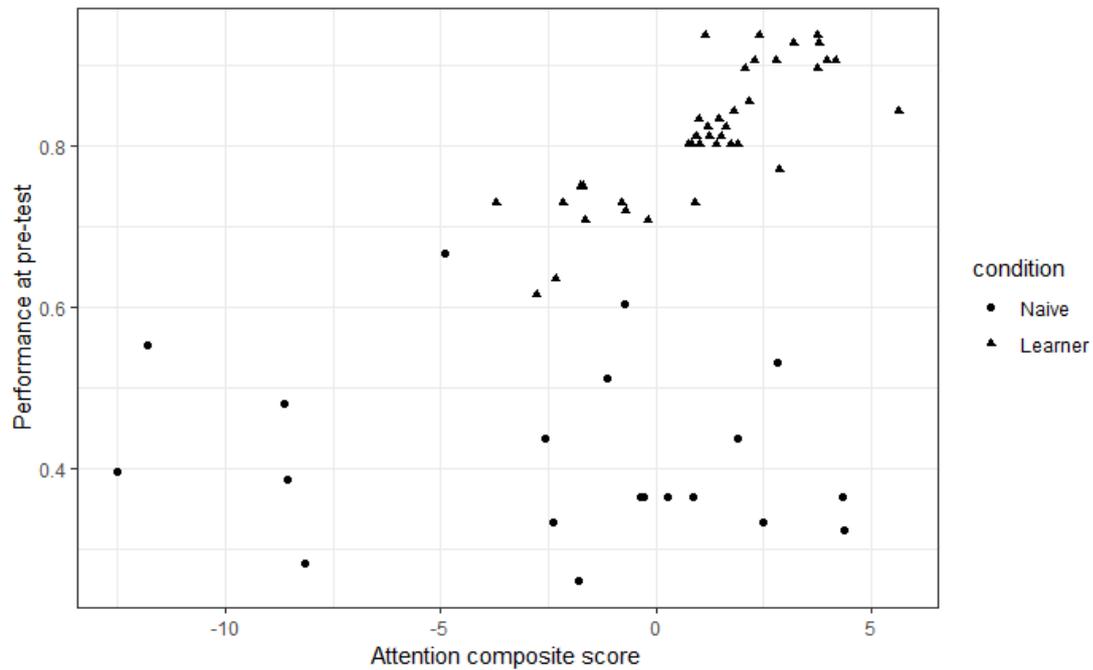


Figure 53 Scatter plot for the Pitch Contour Perception Test with Attention composite score as x-axis and pre-test performance as y-axis.

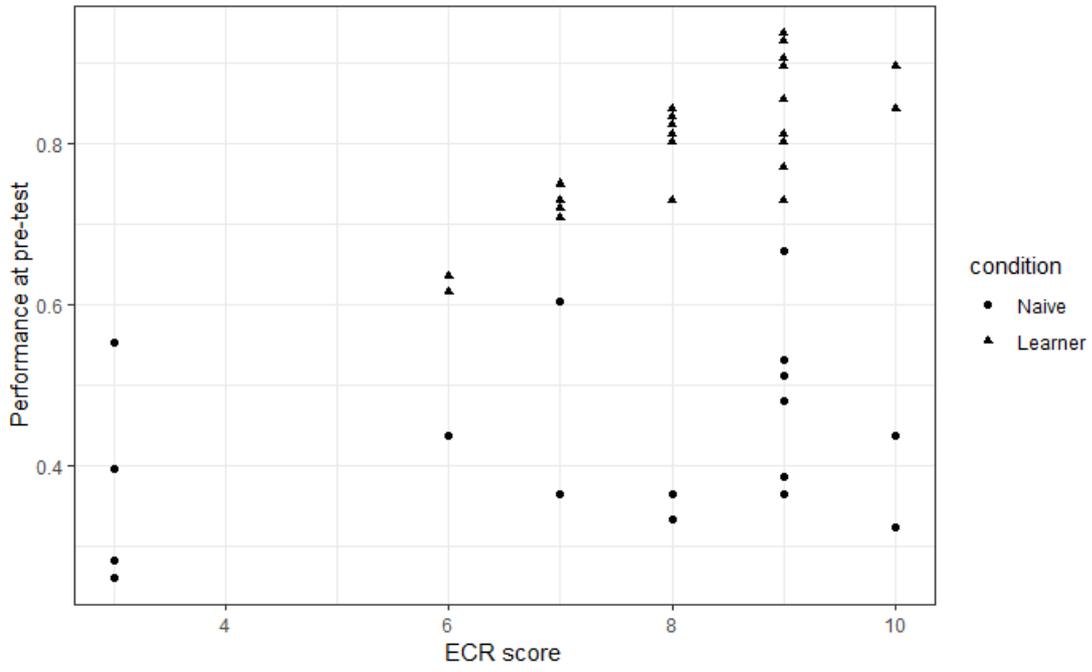


Figure 54 Scatter plot for the Pitch Contour Perception Test with ECR score as x-axis and pre-test performance as y-axis.

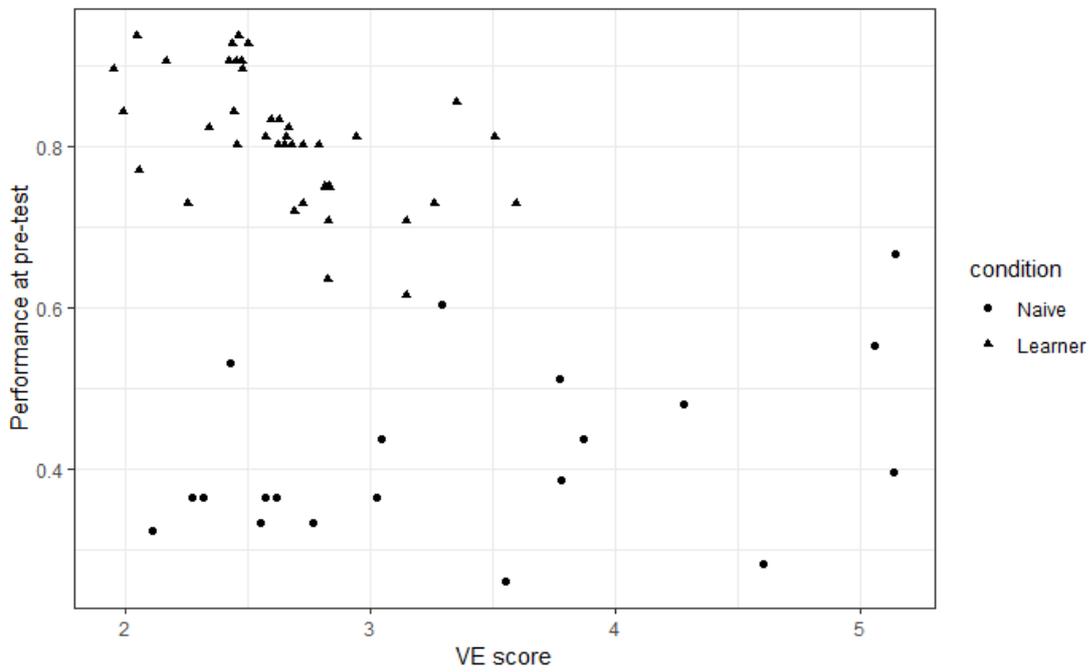


Figure 55 Scatter plot for the Pitch Contour Perception Test with VE score as x-axis and pre-test performance as y-axis.

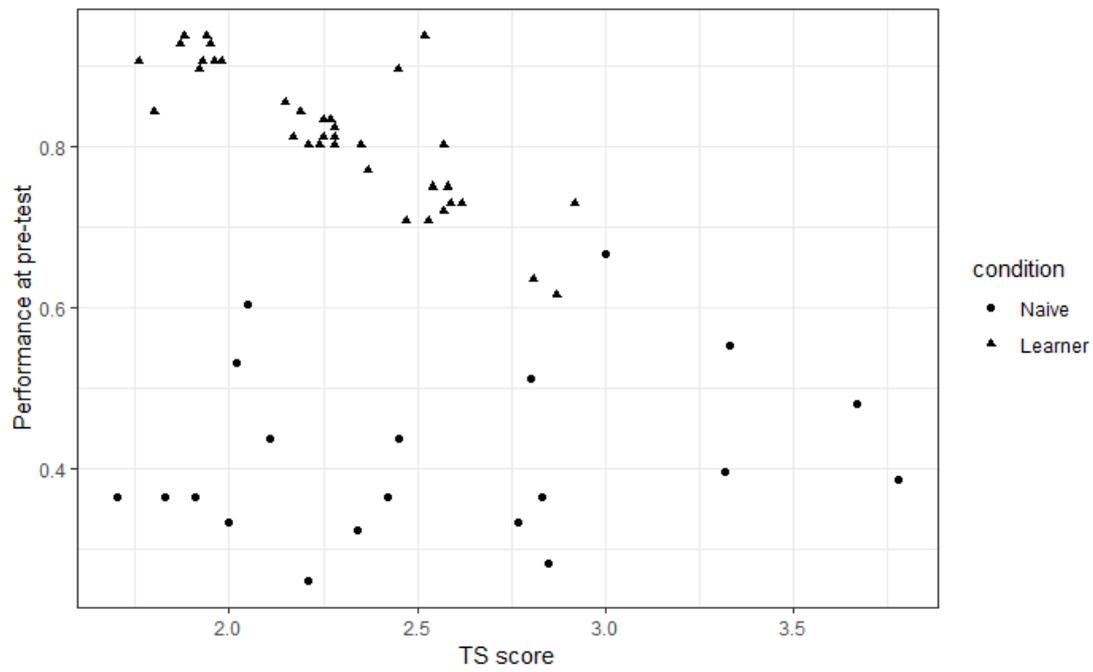


Figure 56 Scatter plot for the Pitch Contour Perception Test with TS score as x-axis and pre-test performance as y-axis.

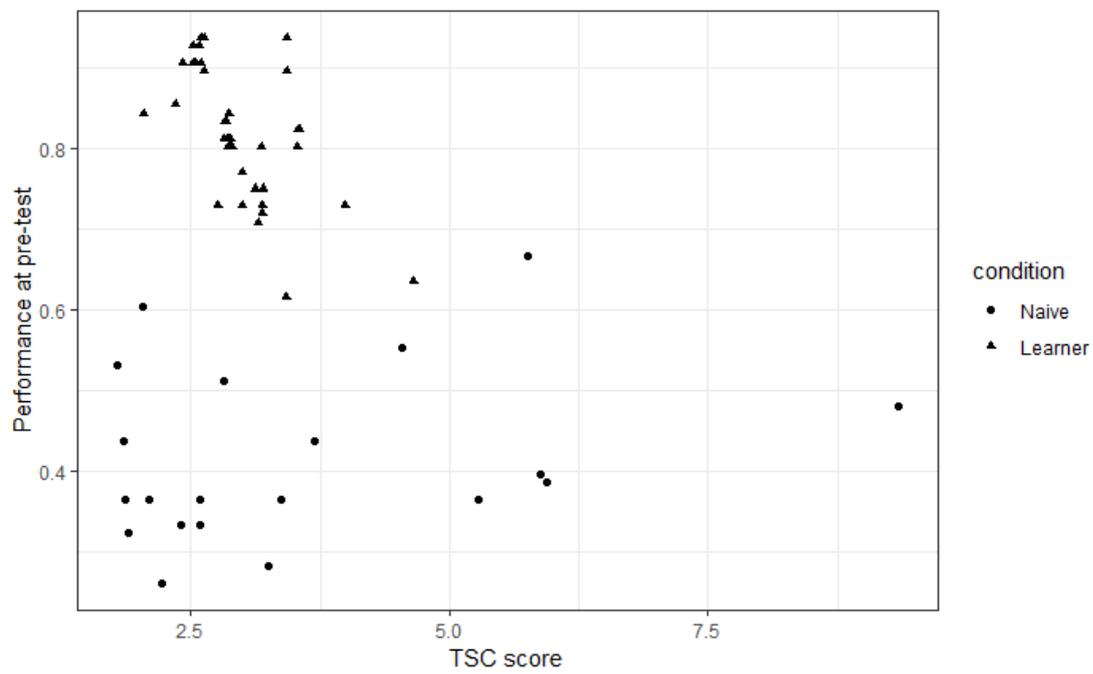


Figure 57 Scatter plot for the Pitch Contour Perception Test with TSC score as x-axis and pre-test performance as y-axis.

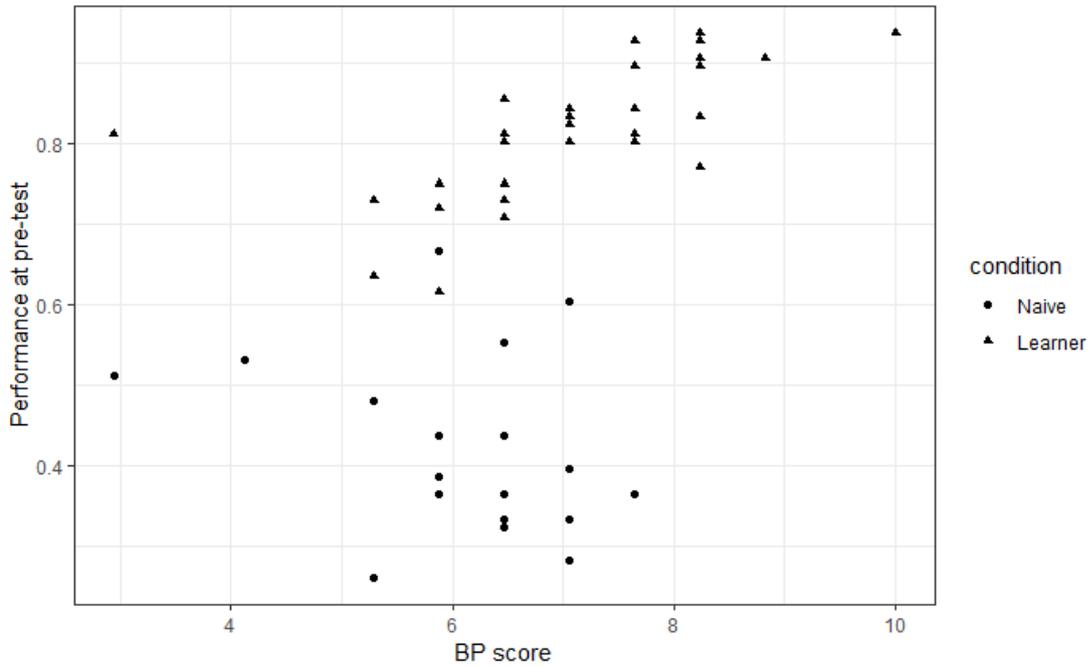


Figure 58 Scatter plot for the Pitch Contour Perception Test with BP score as x-axis and pre-test performance as y-axis.

#### 4.4.2.4.2.2 Hypotheses that *ID measure* predicts pre to post-test improvement, and this differs for the participant groups

Relevant statistics are summarised in Table 25. It should be noted that as the model including interactions with age did not converge, therefore only the main effect of age was included in the model looking at this predictor (as reported in Table 24). For the composite working memory measure, there was near substantial evidence for null, that it did not predict the improvement from pre- to post- test. For all other separate measures, I found evidence for the null, except for *Letter Number Sequencing* where the evidence was ambiguous. Where the BF for the interaction could be computed, I only found ambiguous evidence regarding whether there was a difference between the MLP and the NP groups. For attention measures there was substantial evidence that the composite score was predictive of participants' improvement and a similar pattern was found for *Telephone Search*. However there was substantial evidence that this was modulated by an interaction with participant condition for the composite score and near substantial evidence for this interaction for *Telephone Search*. Breaking down this

composite score (Figure 59) and *Telephone Search* (Figure 61), in both cases there was substantial evidence for the effect in the MLP group only, with evidence for the null for composite score in the NP group, and ambiguous evidence in the NP group for *Telephone Search*. For all other measures the evidence effect that they predicted pre to post improvement was ambiguous and the evidence for the interaction with participant condition was also ambiguous except for the *Elevator Counting with Reversal* (Figure 60) where there was substantial evidence for the interaction. Breaking this down, there was again evidence for an effect only in the MLP, with evidence for the null in the NP group. For musical ability measures, there was substantial evidence that the composite score was predictive of participants' improvements, but the evidence was ambiguous regarding an interaction with participant-condition. Again, the two musical tests patterned differently: For *Melody Memory* there was evidence for the null that it did not predict the improvement, with ambiguous evidence for the interaction with participant-condition. For *Beat perception*, there was strong evidence that it was predictive and that this effect was modulated by an interaction. After breaking down, I found an effect in the NP group only while there was evidence for the null in the MLP group.

Table 25 Regression and Bayesian analysis for Pitch Contour Perception Test, with the effect of *ID measure x test-session* and *ID measure x test-session x participant-condition*

Task	Effect of individual aptitude by test-session (positive $\beta$ indicates larger effect in post-test)						Effect of individual aptitude by test-session by condition (positive $\beta$ indicates larger effect in the MLP group)					
	$\beta$	SE	P	H1	Bayes	Robustness Region	$\beta$	SE	p	H1	Bayes	Robustness Region
Age												
Working memory composite score	0.022	0.04	0.437	0.176	0.333	[0,0.176]	0.07	0.04	0.20	0.099	1.039	[0,0.30]
Digit Span - Forward	0.02	0.04	0.612	0.727	0.087	[0.61,∞]	0.10	0.08	0.22	0.02	1.013	[0,0.51]
Digit Span - Backward	-0.01	0.04	0.711	0.669	0.043	[0.30, ∞]	0.07	0.08	0.371			
Arithmetic	-0.02	0.03	0.531	0.558	0.041	[0.25,∞]	-0.02	0.06	0.775			

Letter – Number Sequencing	0.06	0.03	0.055	0.505	0.725	[0,1.06]	0.08	0.06	0.186	0.059	1.097	[0,0.40]
Attention composite score	<b>0.06</b>	<b>0.04</b>	<b>0.028</b>	<b>0.09</b>	<b>5.086</b>	[0.05,0.15]	<b>0.11</b>	<b>0.04</b>	<b>0.017</b>	<b>0.061</b>	<b>3.946</b>	[0.05,0.20]
Attention composite score <i>NP group only</i>	-0.01	0.02	0.698	0.092	0.189	[0.05,∞]						
Attention composite score <i>MLP group only</i>	<b>0.10</b>	<b>0.04</b>	<b>0.02</b>	<b>0.05</b>	<b>8.496</b>	[0.05,0.05]						
Elevator Counting with Distraction	0.03	0.04	0.386	0.642	0.141	[0.30,∞]	0.11	0.06	0.093	0.0334	1.189	[0,0.76]
Elevator Counting with Reversal*	0.09	0.06	0.092	0.483	0.911	[0,1.31]	<b>0.24</b>	<b>0.09</b>	<b>0.007</b>	<b>0.095</b>	<b>4.579</b>	[0.07,1.10]
Elevator Counting with Reversal <i>NP group only</i>	-0.07	0.05	0.166	0.18	0.118	[0.10,∞]						
Elevator Counting with Reversal <i>MLP group only</i>	<b>0.18</b>	<b>0.08</b>	<b>0.027</b>	<b>0.172</b>	<b>6.285</b>	[0.10,0.56]						
Visual Elevator	-0.21	0.15	0.162	- 1.309	0.547	[0,-2.12]	- 0.26	0.24	0.271	-0.21	0.976	[0,-1.26]
Telephone Search	<b>-0.57</b>	<b>0.20</b>	<b>0.005</b>	- 1.697	<b>13.622</b>	[-0.20,-2.37]	- 0.70	<b>0.32</b>	<b>0.032</b>	<b>-0.564</b>	<b>2.923</b>	[0,>-5]
Telephone Search <i>NP group only</i>	-0.11	0.20	0.576	0.768	0.404	[0,-0.91]						
Telephone Search <i>MLP group only</i>	<b>-0.77</b>	<b>0.26</b>	<b>0.003</b>	<b>0.848</b>	<b>30.619</b>	[-0.25,-3.84]						
Telephone Search while Counting	-0.11	0.12	0.325	- 0.156	1.28	[0,-0.91]	- 0.21	0.18	0.296	-0.114	1.021	[0,-0.96]
Musical ability composite score	<b>0.19</b>	<b>0.05</b>	<b>&lt;0.001</b>	<b>0.397</b>	<b>742.887</b>	[0.05,5]	- 0.17	0.09	0.062	0.19	1.804	[0,1.57]
Beat Perception	<b>0.22</b>	<b>0.05</b>	<b>&lt;0.001</b>	<b>0.625</b>	<b>3602.136</b>	[0.05,5]	- 0.31	<b>0.10</b>	<b>0.001</b>	<b>0.216</b>	<b>29.865</b>	[0.10,1.61]
Beat Perception <i>NP group only</i>	<b>0.42</b>	<b>0.07</b>	<b>&lt;0.001</b>	<b>0.462</b>	<b>&gt;99999</b>	[0.05,>5]						
Beat Perception <i>MLP group only</i>	0.123	0.07	0.123	0.067	1.588	[0,2.07]						
Melody Memory	0.05	0.06	0.348	1.05	0.142	[0.45,∞]	0.08	0.12	0.496	0.054	0.946	[0,0.40]

\*These ID measures were analysed with a larger number of steps (500) to make sure H1 was covered in the Robustness regions calculated.

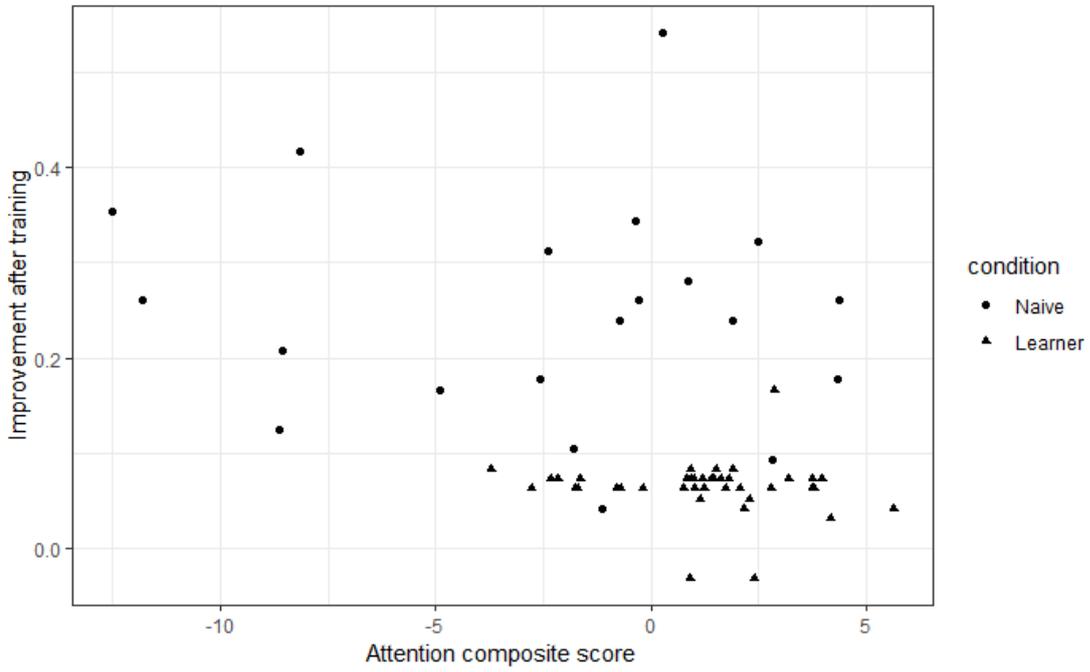


Figure 59 Scatter plot for Pitch Contour Perception Test, with Attention composite score as x-axis and pre/post-test difference as y-axis.

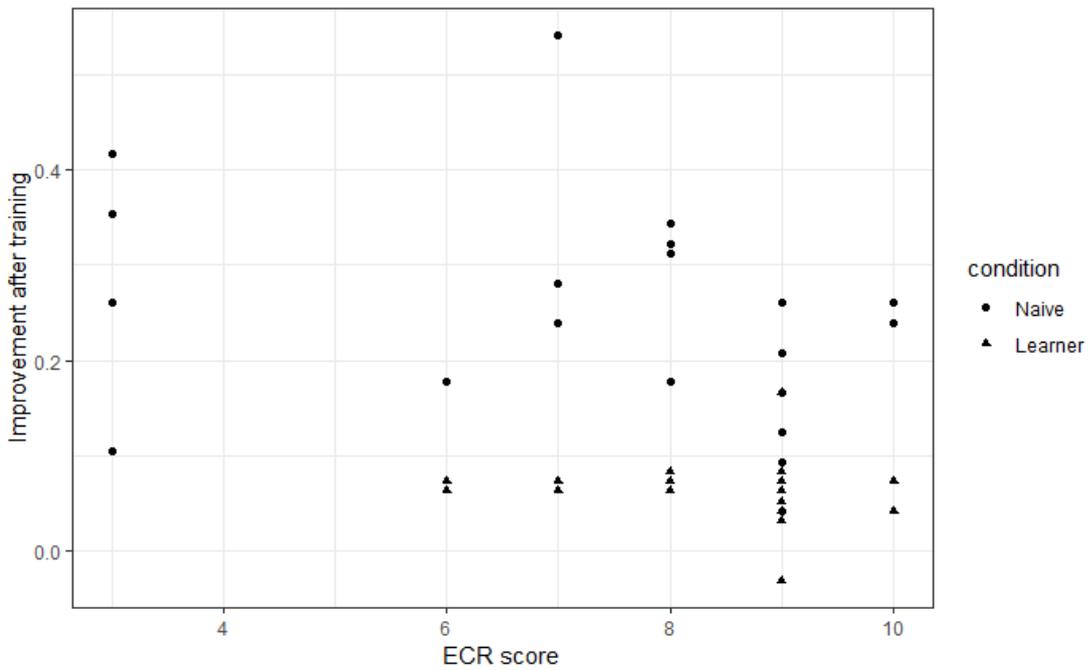


Figure 60 Scatter plot for Pitch Contour Perception Test, with ECR score as x-axis and pre/post-test difference as y-axis.

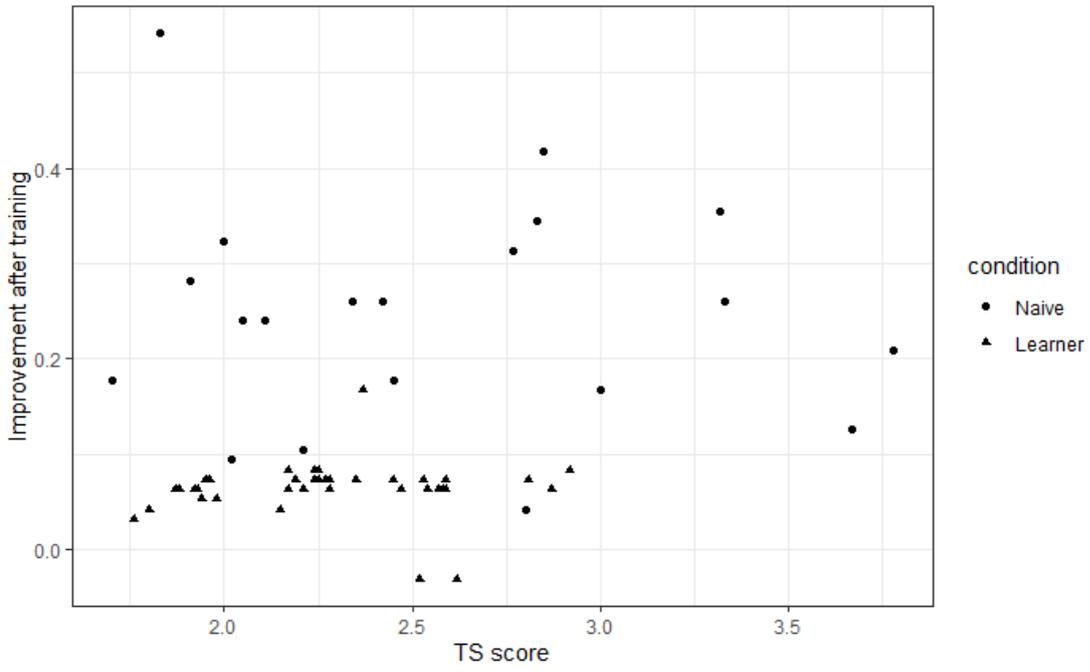


Figure 61 Scatter plot for Pitch Contour Perception Test, with TS score as x-axis and pre/post-test difference as y-axis.

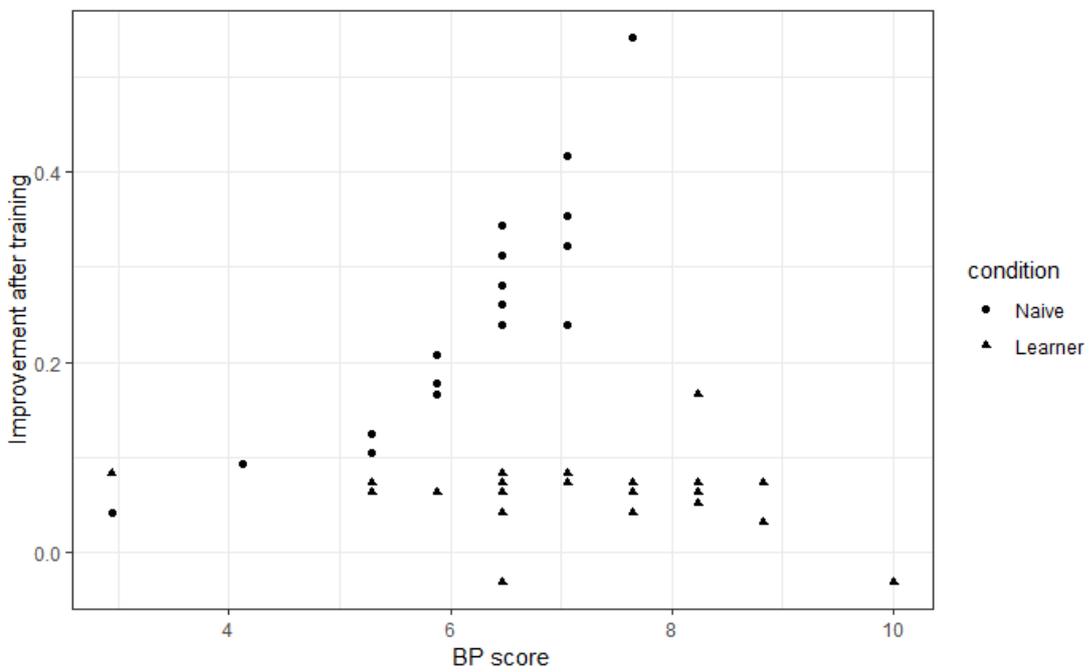


Figure 62 Scatter plot for Pitch Contour Perception Test, with BP score as x-axis and pre/post-test difference as y-axis.

#### 4.4.3 Training Data

A logistic mixed effects models was run with the predicted variable was whether a correct response was given (1/0) on each trial. The predictors were the numeric factor *training-session* (1→4) and the factor *participant-condition* which had two levels (naïve participant; Mandarin learners), both were given a numeric centered coding. The mean accuracy is displayed in Figure 63.

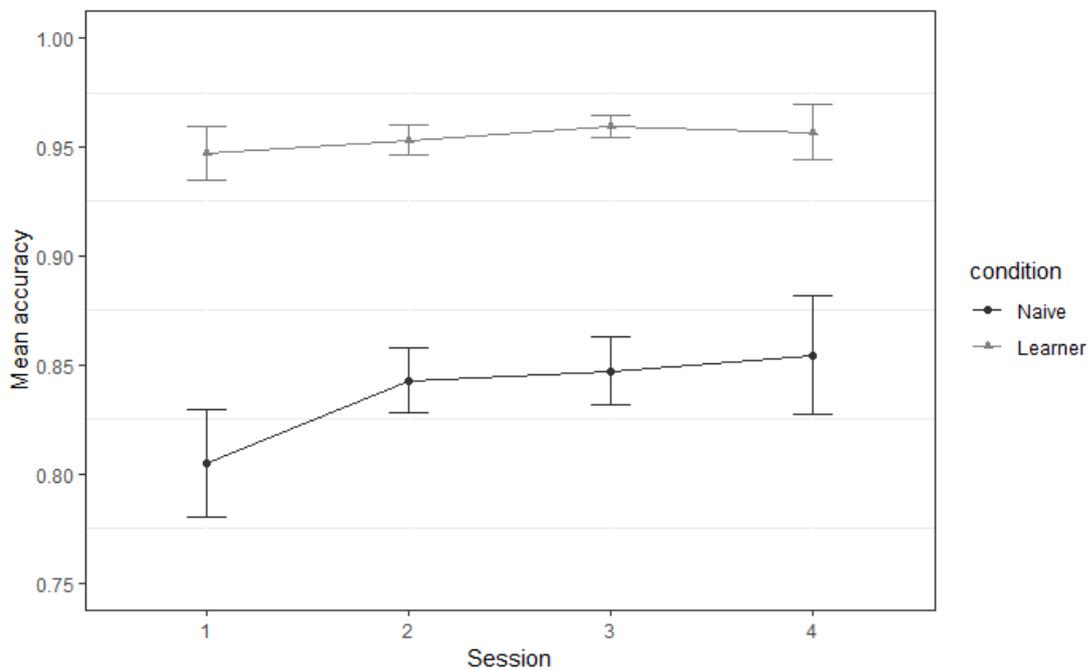


Figure 63 Mean proportion of correct of Naïve participants and Mandarin learners through session 1 to 4.

There was an effect of *training-session* ( $\beta = 0.12$ ,  $SE = 0.04$ ,  $z = 2.89$ ,  $p < .001$ ). Participants' performance increased significantly with training-sessions. Overall, the Mandarin learners performed better than the naïve participants ( $\beta = 1.34$ ,  $SE = 0.28$ ,  $z = 4.73$ ,  $p < .001$ ). There was no interaction ( $\beta = -0.13$ ,  $SE = 0.09$ ,  $z = -1.38$ ,  $p = 0.17$ ).

I did not explore the role of individual difference measures in training since adding the *ID measures* into the relevant model led to convergence issues for most of the measures.

#### 4.5 Discussion

The current study aimed to explore the relationship between individual differences—including measures of Working Memory, Attention and Musical Ability as well as the ability of native English speakers to perceive and produce real Mandarin tones and words before and after high variability computerized phonetic training. It also looked at whether a tone identification task, used in previous studies as a measure of individual difference, was predictive of performance in the current learning tasks. The study looked at this in naïve participants (NP) and participants who were students of Mandarin (MLP).

There were four performance measures, collected pre and post. Two production measures: Pinyin reading tone accuracy – whether participants can correctly produce the tone in the reading task, and pinyin reading pinyin accuracy – whether participants can correctly produce the pinyin in the reading task; Two perception measures: Four Interval Oddity task – whether participants can correctly select the Mandarin word with the different tone out of a choice of four, and the Pitch Contour Perception Test – whether participants can correctly identify the Mandarin tone in the context of six Mandarin vowels.

Examining the results without considering *ID measures*, it can be seen that, in production, as expected the MLP group outperformed the NP group for each measure at pre-test. In terms of pre-to-post performance for both production measures, there was only evidence that naïve participants showed increased performance, however performance of Mandarin learner participants was near ceiling even at pre-test, although they did show significant improvement smaller than naïve participants. For both of the two perception measures, there was evidence that both groups improved after training, but the improvement was more notable for Mandarin learners in the Four Interval Oddity task and for naïve learners in the Pitch

Contour Perception Test. For the data from training itself, both groups improved significantly across training sessions, although learning was more modest in the Mandarin learners, due to ceiling effects. These results support previous research (e.g. Logan et al., 1993, Perrachione et al., 2011) showing that high variability training is effective when learning new phonetic contrasts, including Mandarin tones, and support the findings of Study 1 and Study 2 that this extends to the learning of all four Mandarin tones in the context with all six Mandarin vowels in real Mandarin words. It also demonstrates that HVPT is effective even for the learners who have already acquired some knowledge in that language. Similarly to Study 1 and Study 2, it can be seen that the major improvements in training were made between session 1 and session 2, highlighting the possibility that participants will be able to pick up novel phonetic features relatively quickly. The results also revealed that the effect of HVPT can be extended to new items and words, as shown in previous literature and in Study 1 and 2.

For the analyses with *ID measures*, the 11 measures of Working Memory, Attention and Musical Ability were used as predictors. Composite scores were also computed for Working Memory, Attention and Musical Ability separately, aiming to gain a more robust measure of these general effect. Participants' scores on the PCPT task at pre-test were used as an additional predictor in the models with production measures and Four Interval Oddity task as outcomes. Although not part of the research question, age was also used as a predictor to see if this contributed to differences at pre-test or degree of learning. This is important since our participant samples were not well matched in terms of age, with the MLP group being significantly younger than the NP group. The analyses for Study 3 explored the following three questions about individual differences: firstly, the models examine whether there was a difference between the two participant groups in terms of their scores on the *ID measures* even before training. Findings suggested substantial evidence that the MLP group outperformed the

NP group on the *Pitch Contour Perception Test* and all but one of the other *ID measures* (where the evidence was ambiguous). This is discussed further in Section 4.5.1. Secondly, the analyses also looked at whether these *ID measures* predicted the performance of participants in the outcome measures at *pre-test* and whether the pattern was different between groups. Thirdly, this set of analyses examined whether these *ID measures* predict the improvements of participants from pre-test to post-test and whether the pattern was different for the two groups. For both performance at pre-test and pre-to-post improvement, across all tasks and measures, age was not predictive of pre-test performance or improvement, with evidence for the null in each case (although recall that I could not test the interaction between age and test-session or participant condition for *Pitch Contour Perception Test* outcome measures due to model convergence problems). In general, patterns of which *ID measures* were predictive differed for different outcome measures, and in some cases the patterns were also different between participant groups. This interesting yet complex pattern of results is discussed further in Section 4.5.2. Finally, Section 4.5.3 discusses the possible limitations of the current design and potential future research directions.

#### 4.5.1 *Baseline differences between Naïve Participants and Mandarin Learning Participants*

An advantage for Mandarin learner participants over naïve participants was seen on all but one of individual differences measures (with ambiguous evidence for the attention measure *Telephone Search while Counting*). What could account for these differences between our participant groups? One explanation is that their previous experience learning Mandarin led to increased ability in these other areas. This is undoubtedly the case for the *Pitch Contour Perception Test* which directly measures one's knowledge of Mandarin tones. However, it is less clear that Mandarin learning experience would have led to benefits for the tests of other cognitive abilities. Therefore, I will consider how such an explanation would fit with previous

literature for each individual difference measure in turn. I will then consider more generally the alternative hypothesis that these differences between participants are not due to Mandarin learning experience but due to other differences between the natures of these two groups of participants.

For working memory, there is currently no similar direct comparison between English speakers who are and are not learning Mandarin on these types of measures. There is some evidence for an advantage on digit span task score in native Mandarin speakers compared with native English speakers (Chen, Cowell, Varley and Wang, 2009). However this is most likely due to the fact that the tests were conducted in different languages with different temporal features (Chinese number words have significantly shorter pronunciation duration than English number words, Stigler, Lee & Stevenson, 1986). This explanation is supported by a recent study by Mattys, Baddeley & Trenkic, (2018) who compared the performance of native Mandarin speakers who had learned English with native English speakers using both digit span and word span tasks administered in the two languages. Although Mandarin speakers performed better on both Mandarin digits and words span than English speakers performed on English digits and words span, they did not exhibit any advantage on either English words or numbers over English speakers. This suggests that these differences are due to the use of Mandarin stimuli rather than broader differences in cognitive function due to learning Mandarin, making it less relevant to the current study. There is some evidence that learning Mandarin written words may affect working memory. The logographic system of writing in Mandarin involves separating each character into a series of strokes within a square grid (Tan et al., 2001) and this has been linked to increased spatial memory capacity found in Mandarin speakers (Chen et al., 2009). Perhaps the participants' exposure to Mandarin characters in their course has benefited the capacity of working memory in general. However, it should also be noted that the current study

only employed phonological, not visual working memory tasks, so that any benefit would have had to transfer from visual to phonological working memory.

One of the specific tests of working memory is a test of arithmetic ability. It has been argued that Mandarin is superior to English for processing mathematic problems, and that this may underpin differences in mathematical exam results compared between Mandarin and English speakers (Galligan, 2001). One reason suggested by Han and Ginsburg (2011), is that Mandarin words have higher clarity than English for describing mathematical-related terms as these Mandarin terms involved compound words with a more transparent mapping to the mathematical concept. Han and Ginsburg (2011) also compared U.S. urban junior high school students' mathematical abilities. Mandarin-speaking students were better than English speaking students, even though the tests were administered in English, and their performance was also correlated with Mandarin reading ability providing stronger evidence for a causal relationship between Mandarin learning experience and mathematical skills. This suggests that learning Mandarin, may lead students to be more efficient when solving mathematical problems presented in English, as they've learned a more efficient way of representing those numbers. Given this evidence, it is possible that learning Mandarin as an L2 is beneficial for solving mathematical questions, and thus improves participants' performance on our arithmetic test. However, in the context of current study, although the Mandarin learner participants will have learned Mandarin words for numbers, it is unlikely that they have specifically practiced Math calculations in Mandarin. This makes less likely the reason why the MLP group outperform the NP group in this test.

For attention measures, the MLP group exhibited an advantage on all measures except *Telephone Search while Counting* (where the evidence was ambiguous). Again, there is no previous direct comparison of these measures between English speakers who do and don't learn

Mandarin. More broadly, however, there is the notion of a general “bilingual advantage” in attention, which has been shown in several languages (e.g. French-English: Morton & Harper, 2007; German-French: Reder, Marec-Breton, Gombert, & Demont, 2013), including English-Mandarin bilinguals as shown in a series of work (e.g. Bialystok, 1999; Bialystok & Martin, 2004). These studies demonstrated a difference between monolinguals and bilinguals such that bilinguals have cognitive advantages especially in the control of attention. Most studies focus on early bilinguals (Bialystok, Craik, Klein & Viswanathan, 2004) so there is a lack of evidence regarding whether the attention skills of late learners can still benefit from learning a second language. However a recent study (Bak, Vega-Mendoza & Sorace, 2014) compared early childhood bilinguals (exposed to a second language by age 4), late childhood bilinguals (exposed to a second language between 4 and 15), early adult bilinguals (exposed to a second language between 15 and 19) and monolingual English speakers on their performances on Test of Everyday Attention. Results suggested that even early adulthood bilinguals demonstrated an advantage over monolingual speakers on selective attention but only non-significant advantage was found on attention switching measures. The main explanation in the literature for why there is an advantage from learning a second language is the idea that bilinguals need to constantly suppress the activation of the non-target language, resulting in enhanced executive function (e.g. Ong, Sewell, Weekes, McKague & Abutalebi, 2017). However, in recent years, the bilingual advantage has been challenged. For example, Paap and Greenberg (2013) ran a series of experiments including anti-saccade task, Simon task, flanker task and colour-shape switching task which are all well-used tasks measuring central executive function. The only difference between groups was a bilingual *disadvantage* in one task, and there was also no consistent pattern across different tasks which are meant to measure similar aspects of attention (e.g. inhibitory control in Simon and flanker tasks). A systematic review was run by Van den Noort et al. (2019) looking at 56 relevant study between 2004 and 2018. Only 54.3% of these

studies reported an advantage of bilinguals on attention. They concluded that while these studies largely depend on methodology (e.g. the selection of tasks, the recruitment of participants), there may be a bias towards both for and against bilingual advantage (i.e. researchers tended to report only significant results that there is a bilingual advantage and potentially ignore null results). Thus in the current experiment it is possible the advantage in attention measures for the MLP group is due to benefits from learning Mandarin, however, this should be interpreted with caution given the conflicting findings in the literature.

For musical ability measures, a clear advantage for Mandarin learners was revealed on both tests on *Beat Perception* and *Melody Memory*. Better musical ability is linked to several aspects of linguistic processing involved in Mandarin such as pitch discrimination (Magne et al., 2006) and lexical tone discrimination (Delogu et al., 2006), which is likely due to the fact that Mandarin tones share similar pitch pattern mechanism as music tones. This is reviewed more extensively in section 4.1.3.1. There is no direct evidence that Mandarin learners develop their musical ability via learning Mandarin tones. However there is some more general evidence that second language learning could increase musical ability. A recent study (Bhatara, Yeung & Nazzi, 2015) has suggested a relationship between foreign languages learning experience and rhythm perception. They recruited 147 monolingual European French speakers. Using the Musical Ear Test (Wallentin et al., 2010), they measured their performance on rhythm and melody perception. Longer foreign language experience (measured by year of involvement) was highly correlated with better rhythm perception but not melody perception. For Mandarin in particular, it is possible that participants benefit from exposure to a language with different types of rhythmic class distinctions: English is considered to be “stress-timed”, so that that listeners perceive the language based on the segmentation by stress (Pike, 1945), while Mandarin is seen as “syllable-timed” so that the segmentation is based on syllables which

takes approximately similar time to pronounce (Lin & Wang, 2007). Learning new multiple rhythm patterns may make participants more equipped when learning music patterns.

To summarise, the discussion above suggests it is possible that learning Mandarin has benefited the participants in the current design, particularly in the areas of Attention and Musical Ability. However, no conclusion regarding causal relationship can be drawn based on the current results. It could be that participants who have better Working Memory, Attention and Musical Ability are more likely to choose to study a second language such as Mandarin. It is also possible that other demographic differences might explain the different cognitive profile of the two groups. For example, the MLP group is younger than the NP group. It is also worth noting that the MLP group are all current university students, while the majority of the NP group have already graduated from university. It may be that those who are currently students are simply more in the habit of taking tests and more motivated.

#### 4.5.2 *Relationship between measures of individual differences and performance on the pre-post tests*

##### 4.5.2.1 *Production tasks*

The findings with the production tasks were limited in the current design. At pre-test, the only task found to be predictive was *Digit Span Backward*. This predicted performance in Pinyin accuracy, although this was modulated by an interaction with *participant-condition*, which broke down to show that there was only evidence for this effect in the NP group (ambiguous evidence for the MLP group). This pattern was reflected at the WM composite score, however evidence for other individual WM measures was either ambiguous (*Digit Span Forward*) or for the null (*Arithmetic & Letter Number Sequencing*). Otherwise: For tone accuracy no *ID measure*, including the *Pitch Contour Perception Test*, predicted performance, with evidence for the null in each case; For Pinyin accuracy there was evidence for the null for

the *Pitch Contour Perception Test* and for the Attention composite score, with either evidence for the null or ambiguous evidence for each of the subtests and for Musical Ability measure there was evidence for the null for the composite score as well as the two individual scores. In all cases where there was no evidence for an overall predictive relationship with an *ID measure*, the evidence that the effect was modulated by group was (where it could be computed) ambiguous.

Turning to improvement from pre- to post-test, the only place where there was evidence for a predictive relationship was the strong evidence that DSB predicted improvement in tone accuracy, although there was an interaction with group and this only held for the NP group, with evidence for the null for the MLP group. This was also reflected in the Working Memory composite score (although here the evidence for the MLP group was ambiguous). Otherwise, there was a mixture of null/ambiguous evidence: For tone accuracy, evidence for the *Pitch Contour Perception Test* was ambiguous; For Working Memory measures other than *Digit Span Backward* (*Digit Span Forward*, *Arithmetic* & *Letter Number Sequencing*) evidence was ambiguous; For Attention measures, there was evidence for the null for the composite score and for two separate subtests (*Elevator Counting with Distraction* & *Elevator Counting with Reversal*) while for the other two (*Telephone Search* & *Telephone Search while Counting*) the evidence was ambiguous; For Musical Ability, there was ambiguous evidence for the composite score and for BP, and evidence for the null for *Melody Memory*. For Pinyin accuracy, for all ID measures (and composite scores) there was evidence for the null except for *Pitch Contour Perception Test* and the attention measures *Elevator Counting with Reversal*, *Visual Elevator* and *Telephone Search while Counting*, where the evidence was ambiguous. For both tone accuracy and Pinyin accuracy, again in all cases where there was no evidence for an

overall predictive relationship with an *ID measure*, the evidence that the effect was modulated by group was ambiguous (where it could be computed)

In interpreting these patterns, it is important to note that there may be different explanations for the results for the two participant groups. For the MLP group, it should be acknowledged that although a clear null results were found in many cases, this is likely due to ceiling effects which were present even at pre-test (Tone accuracy: 91%; Pinyin accuracy: 92%). Thus it is not too surprising that there was no relationship between these tasks and *ID measures*. Further interpretation will therefore focus on understanding the results from naïve individuals.

For the NP group, why might *Digit Span Backward* predict both pre-test performance in the pinyin accuracy and pre- to- post improvement in tone accuracy? Starting with pinyin accuracy, note that at pre-test, as these participants had no prior knowledge of Mandarin tone or Pinyin, this task can only be treated as a non-word reading task. This may explain why their performance of Pinyin was predicted by *Digit Span Backward*, as non-word reading task performance has been found to be correlated with WM measures. Carretti, Borella, Cornoldi & De Beni (2009) conducted a meta-analysis using 18 studies from 1989 to 2006. They found a strong correlation between non-word reading performance and working memory measured with digit span forward (4 studies) in children and adults with specific reading comprehension difficulties. Exactly why it is *Digit Span Backward* that is predictive in the current design, is unclear, as the meta-analysis also revealed WM measured by digit span forward also correlated with non-word reading performance, though note that the relationship for the other three Working Memory measures is ambiguous, so strong conclusions cannot be drawn about these tests.

Turning to tone learning, in the literature, there is no direct study on the effect of working memory on learning to produce Mandarin as a second language. Some evidence (O'brien, et. al., 2006) suggests that phonological working memory predicts the richness of vocabulary use for L2 learners in the early stages. In addition, recall from Section 4.1.1.4 that previous studies have also found that WM can predict the rate at which children acquire L1 vocabulary (Gathercole et al., 1992) as well as adults' learning rate of L2 vocabulary and grammar (Bergsleithner, 2011; Cheung, 1996; Service & Kohonen 1995; Speciale, et al., 2004; Weissheimer & Mota, 2011). However the current study did not look at vocabulary learning specifically but focused on phonological accuracy in production. It is possible that the role of working memory in vocabulary learning is at least partially relevant to developing phonological accuracy. The reason why the current analysis only found working memory measures predicted the improvement of tone learning, and not the segmental phonology (with evidence for the null), might be that tone is the new phonetic feature and is targeted by the training. Thus participants' cognitive resources may not be evenly distributed on tones and segments. Recall that Ou and Law (2017) actually found auditory working memory measured by digit span backward did *not* predict native Cantonese speakers' tonal production performance. However, it should be noted that they did not use Bayesian analysis so it is not clear whether they actually found evidence for the null and they did find visual working memory measures predicted participants' performance. They attributed the lack of correlation with auditory working memory measure to the nature of measures. The production performance was measured by voice onset time and visual working memory also used RT as the measure. They concluded that RT measures may better capture the changes in the current production task. Again, it is unclear why it is specifically digit span backwards that is predictive in the current design, however given that evidence for the other working memory measures was ambiguous I shall not over interpret this. It is also unclear why this measure is predictive of

tone learning for production measures but *not* in either of perception measures (with evidence for the null in this case) (see Section 4.5.2.2.2), a point to which I return below.

The current study did not find any Attention measure predicting the pre-test performance or the improvement from pre- to post-training, and this was found for both tone and Pinyin measures, with evidence for the null results at least for the composite scores. As suggested above, for the MLP group this could be due to ceiling effects, but this isn't the case for the NP group who did show considerable improvement in performance from pre- to post-test. Again, there is no direct research specifically on L2 Mandarin production and attention. As discussed in Section 4.1.2.1, it has been suggested in previous literature that deficits in attention ability may impair language abilities such as reading (Lallier et al., 2010) and people who perform better in attention tests are better at discriminating second languages phonetic contrasts (e.g. Díaz, 2008; Hazan and Kim 2010). The explanation might be that people with higher attention measures can tend to new words quicker and focus on them for longer period of time. However these tests are generally perception tests- the current study may suggest that these abilities may not transfer to producing Mandarin tones and Pinyin.

For musical ability, there was also no evidence that either measure predicted pre-test performance, with evidence for the null for both cases, or improvement after training, with evidence for the null for *Melody Memory*. In contrast in the literature, there is evidence that musical ability has a positive relationship with L2 production. For instance, recall from Section 4.1.3.1, Slevc and Miyake (2006) reported that musical ability is a strong predictor of L2 productive skills, even after controlling of other factors such as age of L2 immersion, patterns of language use and exposure, and phonological short-term memory. There is also evidence that native English speakers who are musicians are better at producing Mandarin tones than non-musicians (Gottfried, Staby & Ziemer; 2004). Most relevant to the current study, recall

that Li and DeKeyser (2017) *did* find a relationship between musical ability and tone production in their training study. Moreover their measure of musical ability contained a test rather similar to the current *Melody Memory* test, although the current version is harder. This makes the fact that there is a null result for *Melody Memory* particularly surprising (note that there was ambiguous evidence *Beat perception*, so that this result can't be interpreted). One possibility is that this failure is due to the fact that the current design only trained the participants in perceptual tasks, not production tasks. Although improvement in perceptive skills was transferrable to productive skills- as shown in NPs' improvement from pre- to post-test in production (and seen in Study 1 and 2 as well), Li and DeKeyser (2014) found that benefits in production were far greater after targeted production training than after targeted perception training.

Finally performance in the *Pitch Contour Perception Test* at pre-test was not predictive of performance at pre-test for either of the production measures, with evidence for the null. This contrasts with Study 2 where it was found that *Pitch Contour Perception Test* predicted the overall performance of naïve participants on tone accuracy and pinyin accuracy (with marginally significant results). For the MLP group, the null here could again result from the ceiling effects found in production. For the NP group, the lack of effect for *Pitch Contour Perception Test* may again be due to the use of a reading task in the current experiment. Reading pinyin and diacritics (Study 3) may be fundamentally different from producing Mandarin words and tones from memory. As in Study1 and Study2, for both production measures, the evidence that *Pitch Contour Perception Test* predicted pre- to post-improvement was ambiguous.

#### 4.5.2.2 Perception tasks

The two perception tests, Four Interval Oddity task and Pitch Contour Perception tests yielded relatively similar results. However, the findings for each of the different types of ID measures was relatively complicated and so this section considers findings for each type of cognitive measure separately.

##### 4.5.2.2.1 Perception tasks and PCPT

Participants' scores in the *Pitch Contour Perception Test* at pre-test were used as a predictor in the Four Interval Oddity task. For both pre-test performance and pre-to-post improvement, there was strong evidence that *Pitch Contour Perception Test* predicted performance, however in each case there was an interaction with group, which broke down to show that was only evidence for the MLP group, while for the NP group there was evidence for the null at pre-test and ambiguous evidence for post-training improvement. The fact that this relationship was only seen at pre-test for the MLP group was surprising given that in Study 2 (Section 3.3.6), the *Pitch Contour Perception Test* was found to be predictive of the performance of naive native English speakers in Three Interval Oddity task. One explanation is that the current Four Interval Oddity task was harder due to involving all four Mandarin tones in each trial (Study 2 pre to post: 58%-65%; Study 3: 53%-60%). This may somehow obstruct the relationship with the ability to identify tones for the NP group. Importantly, it was found, for the first time, that the *Pitch Contour Perception Test* was predictive of improvement after training, here at least for the MLP group (with ambiguous results for the NP group). This indicates that at least for the Mandarin learners, those who started with better ability to identify the tones, benefited more from HVPT, which corresponds to the findings in the studies by Perrachione et al. (2011) and Sadakta and McQueen (2014).

#### 4.5.2.2.2 Perception tasks with Working Memory measures

For performance at pre-test, the results were very similar across the Four Interval Oddity task and the Pitch Contour Perception Test. There was strong evidence for an effect of the *Letter Number Sequencing* score in the MLP group only, with evidence for the null in the NP group, and this pattern was reflected in the Working Memory composite score. For other Working Memory subtests, the evidence was mostly ambiguous except there was evidence for the null for *Digit Span Forward* predicting pre-test performance on Four Interval Oddity task. For group differences on these measures, the evidence was ambiguous.

Turning to pre-to-post improvement, different patterns were found for the two tasks. For Four Interval Oddity, there was substantial evidence that *Digit Span Forward* was a predictor for the NP group only, with ambiguous evidence for the MLP group. The same pattern was shown in the Working Memory composite score. The evidence for other Working Memory scores was either ambiguous (*Letter Number Sequencing & Arithmetic*) or there was substantial evidence for the null (*Digit Span Backward*). In contrast, for the Pitch Contour Perception Test, there was no evidence that any Working Memory measure predicted the improvement. Although there was evidence for the null for three out of four measures (*Digit Span Forward, Digit Span Backward, Arithmetic*), the evidence for the composite score was ambiguous. Wherever there was evidence for the null or ambiguous evidence for an overall predictive relationship, the evidence that this differed between groups was ambiguous.

How do these results sit with previous literature? Recall there is evidence in the literature that digit span tasks predict tone discrimination in non-linguistic tasks (George & Coach, 2011; Strait et al., 2010). However, the few studies looking at tone discrimination in a linguistic context have failed to find a relationship between some working memory measures

and discriminating tone: In both Ou et al. (2015) and Ou and Law (2017), they did not find a connection between *Digit Span Backward* performance and the ability of discriminating Cantonese tones in native speakers. The authors attribute the null results to the fact that digit span tasks did not use RT measure while the performance measures in their experiments were RT based, believing that for native speakers, a speeded component may be more sensitive in capturing the relationship between one's efficiency of allocating memory resources. However recall that Chandrasekaran et al. (2010) also did not find a relationship between learning of tones in their training study, and working memory measured using *Digit Span Backward* and *Letter Number Sequencing* tasks. However, a limitation of that study was that their participants were classified as good/bad learners of Mandarin tones, rather than using a continuous measure of performance, which may be less sensitive. They also only had 16 participants which may be too small to find evidence of individual differences for cognitive functions. Finally, they only used naïve learners, which may explain why they did not find a relationship with *Letter Number Sequencing* test, in contrast to the present study. Again, this study did not use Bayesian analyses so it is important not to over-interpret their null findings.

In the current study, there was no evidence that either of the digit span tasks, *Digit Span Forward* and *Digit Span Backward*, predicted pre-test performance for either the Four Interval Oddity task or the Pitch Contour Perception Test. In the case of *Digit Span Forward*, there was actually evidence for the null for the Four Interval Oddity task. Thus, tentatively, digit span tasks may not be well suited to capturing baseline measures of lexical tone discrimination, consistent with the literature discussed above. In terms of improvement following training, interestingly, for the Four Interval Oddity task, *Digit Span Forward* was found to be predictive of the improvement in the NP group, although *Digit Span Backward* did not predict improvement (with evidence for the null), and neither *Digit Span Forward* nor *Digit Span*

*Backward* was predictive in the Pitch Contour Perception Test (with evidence for the null). This pattern is therefore a little hard to interpret and may reflect Type 1 error (note that *Digit Span Forward* predicting the improvement of the NP group in the Four Interval Oddity task has a BF of 11, which, although constituting substantial evidence, is a much lower level of evidence than many of the BFs reported in this study), although it is somewhat consistent with the finding that *Digit Span Backward* predicted improvement for the NP group's production of tones.

In contrast to the relatively weak evidence for digit span, there is very strong evidence (BF's > 6000) that *Letter Number Sequencing* is predictive of pre-test performance in both the Four Interval Oddity and Pitch Contour Perception Test, although only in the MLP group (with evidence for the null in the NP group). Compared with digit span tasks, *Letter Number Sequencing* is a more complicated task which involves the memorisation of different types of stimuli (letter & number). It is believed to be highly correlated with other traditional working memory measures such as digit span, but to also reflect participants' attention ability (Crowe, 2000). Note that the effect here corresponds with the very strong evidence for a role of attention in the MLP group for these tests, which is discussed in more details in the following section.

#### 4.5.2.2.3 Perception tasks with Attention measures

At pre-test, a similar pattern was found across the two performance measures: There was strong evidence for an effect in the Attention composite score and in four of the five subtests (*Elevator Counting with Reversal*, *Visual Elevator*, *Telephone Search & Telephone Search while Counting*). However, each of these measures also demonstrated a group difference, with strong evidence in the MLP group and ambiguous evidence for the NP group in the Four Interval Oddity task (except for *Elevator Counting with Reversal* where there was

evidence for the null) and evidence for the null in the NP group for the Pitch Contour Perception Test. For *Elevator Counting with Distraction*, in both outcome measures, the evidence was ambiguous. Wherever there was no evidence for an overall predictive relationship, the evidence that this differed between groups was ambiguous.

Analyses for pre-to-post improvement also yielded similar results across the two perception tests: There was strong evidence that improvement was predicted by the attention composite score in both cases, and in three subtests (*Elevator Counting with Distraction*, *Elevator Counting with Reversal* and *Telephone Search*) for the Four Interval Oddity task and two subtests (*Telephone Search & Telephone Search while Counting*) for Pitch Contour Perception Test. In every case but one, these effects were modulated by an interaction with group which broke down to show that the attention score only predicted the learning of the MLP group, with evidence for the null in the NP group. (The exception was for *Elevator Counting with Distraction* predicting improvement in Four Interval Oddity task, where the evidence for the interaction was ambiguous). For the remaining measures, the evidence that they predicted improvement was either ambiguous (for *Visual Elevator* and *Telephone Search while Counting* predicting improvement in Four Interval Oddity task and Pitch Contour Perception Test) or evidence for the null (for *Elevator Counting with Distraction* predicting improvement in Pitch Contour Perception Test). There were also two occasions where the evidence for the *ID measure* predicting overall improvement was ambiguous but there was evidence that this differed for the participant groups (an interaction with *participant-condition*): For *Telephone Search* in the Four Interval Oddity task, where the interaction broke down to show that there was evidence for null for the NP group but the evidence was ambiguous for the MLP group (partially fitting with the pattern found for other *ID measures*) and for *Elevator Counting with Reversal* in the Pitch Contour Perception Test, where this *ID measure* predicted

the learning of the MLP group but there was evidence for the null in the NP group (fully fitting the pattern with the other *ID measures*). Otherwise, wherever there was evidence for the null or ambiguous evidence for an overall predictive relationship, the evidence that this differed between groups was ambiguous.

Overall, the current analyses provided substantial (and in some cases extremely strong) evidence for a relationship between tone identification and categorical discrimination abilities and Attention, using a variety of measures. Moreover, attention was not only predictive of baseline performance, but also of the extent of improvement on the task following the high variability perceptual training. However, these effects were found to hold only for the MLP group, with a clear null result for the NP group in most cases. This role for attention in tone learning is consistent with what been reported in the literature, particularly for the measures of attention switching (*Elevator Counting with Reversal* and *Visual Elevator*, see Section 4.2.1). Recall from Section 4.1.2 that there is evidence for a relationship between this ability and the tone discrimination in dyslexic children and adults (Facoetti et al., 2010; Hari & Renvall, 2001; Lallier et al., 2010; Tallal & Piercy, 1973) and that the studies by Ou et al. (2015) and Ou and Law (2017) also showed a direct link between auditory attentional switching (*Visual Elevator* response speed specifically) and discrimination of Cantonese tones by native speakers. The current study extends this finding to show that this ability also predicts participants' ability to benefit from training on lexical tones. Along with the current study, the findings suggest that attention shifting contributes to the perception of tonal contrasts. The current study is the first to show that attention shifting is also predictive of the ability to learn lexical tones from HVPT.

These effects of attention have been explained in several ways in the literature. First of all, according to the psycholinguistic grain size theory (Ziegler & Goswami, 2005), since each language has its own phonology, resulting in different sizes of phonological and lexical

representations, if an individual needs to learn a language, then s/he needs to correctly decode these phonological representations. Attention (especially attention shifting) is an important part of this, allowing efficiently processing and segmenting acoustic features. When learning Mandarin, tonal contrasts are a new type of phonological representation that learners need to form. Thus, given a similar amount of exposure, participants with higher attention ability will be able to form more reliable phonological representation and learn novel phonological features better. Secondly, the A2D (attention-to-dimension) theory (Nosofsky, 1986) suggested that individuals have limited perceptual dimensions. Learning new phonetic contrasts is actually reorganising the weight assigned to different sensory dimensions. In the current case, in training the available cues in each trial are constantly changing (e.g. the tone contrasts, the Mandarin word accompanied with the tone) and participants may be switching attention between different acoustic features to search for reliable cues. Thus, people with high attention ability should naturally be better at finding the most useful cues and picking up the tonal differences.

In the current study, *selective* attention (as represented by TS and TSC results) was also found to be a strong predictor for the Mandarin Learner Participants. This seems to be the first time that this has been found in the language learning literature, although selective attention is believed to contribute to perceptual learning in general. Early research (e.g. Trahiotis, Bernstein, Buell & Spektor, 1990; Tallal et al., 1996) suggested that provided training materials are of appropriate difficulty, participants can be trained to selectively attend to the task-related stimuli and thus detect or distinguish sensory stimuli. They found that this training paradigm is most effective when the difficulty of the materials varies from easy to hard (to be adaptive) and no training effect will be found if the materials are too difficult. However, Amitay, Irwin and Moore (2006) did *not* find this same benefit of adaptive training in a task in which

participants needed to discriminate tones and select the stimuli with different frequencies out of a choice of three (a task similar to the current Four Interval Oddity task). Surprisingly, even one group who received identical training materials (i.e. they kept hearing 1k HZ stimuli and there was actually no “different” item in each trial, so they were basically randomly guessing) showed increased thresholds for identifying frequency differences around 1k HZ, and their magnitude of improvement was no different compared with groups who had received the adaptive training method. This may imply that the training helped participants to form an “anchor” in phonological representation thus they will be able to identify the sound which is not the anchor or discriminate sounds around the anchor. Following the A2D theory mentioned above (Nosofsky, 1986), this series of studies highlighted that even with limited exposure, humans are capable of forming selective attention towards certain pitch height patterns. In the current study, it can be seen that at least for the Mandarin learning participants, those with better selective attention have formed better phonological representations of Mandarin tones when they start the experiment, and benefit more from the same amount of exposure in training.

Turning to the naïve participants, their performance did not appear to be related to these attention measures. In all the cases where there was a group difference, there was evidence for the null in the NP group. This suggests that attention switching skills and the ability to selectively attend to certain stimuli may be less important in the beginning stages of learning. Recall from Section 4.1.2.1 that Zou et al. (2017) has suggested intermediate Mandarin learners allocate attentional resources differently compared with beginners and naïve participants. For intermediate learners, they automatically attend to both tone and segments in a more integrated way, so they need to inhibit the activation of segments if tone is the learning target. Thus, for the MLP group, attention switching skills and selective attention may be particularly useful if they are deliberately focusing on different aspects of the stimuli in different trials. For the NP

group, such skills may be less relevant. It could be the case that even if a naïve participant can allocate more attentional resource to a certain word than another participant, it is not enough for them to form a reliable phonological representation. Compared with the MLP group who had a clear knowledge of the Mandarin phonology, naïve participants may not be able to form accurate, categorised representations of Mandarin tones (i.e. they can notice a differences between tones but do not know for sure how these fall into categories). Thus, the ability to identify the tones in terms of the diacritics may be more relevant, and this maybe be related more to musical ability than attention, as discussed below.

#### 4.5.2.2.4 Perception tasks with Musical Ability measures

At pre-test, the two performance measures showed slightly different patterns. For both Four Interval Oddity and Pitch Contour Perception Test, there was evidence for an effect of *Beat perception* but evidence for the null for *Melody Memory*. However, in Pitch Contour Perception Test, the effect of *Beat perception* was found only on the MLP group but not the NP group, with evidence for the null. The Musical Ability composite shared the same pattern as *Beat perception*. Pre-to-post improvement also exhibited different patterns for the two performance measures. For Four Interval Oddity, *Melody Memory* predicted learning only for the NP group, while there was evidence for the null for the MLP group. For Pitch Contour Perception Test, *Beat perception* predicted learning but again, only for the NP group, although the results for the MLP group was ambiguous. There was ambiguous evidence for *Beat perception* predicting improvements in Four Interval Oddity but evidence for the null for *Melody Memory* predicting improvements in Pitch Contour Perception Test (in both cases the evidence for a difference between groups was ambiguous). The composite score again patterned with BP in each case.

Although this precise pattern of the results, i.e. the difference between two Musical Ability measures for the two outcome measures, is hard to interpret, it is not surprising that musical ability predicted tone learning. Tonal languages share a lot of similarity with music pitch patterns including the fact that individuals need to differentiate pitch heights. Recall from the Section 4.1.3.1 that Delogu, et al., (2010) found evidence that melodic proficiency and music expertise predicted tonal identification for Mandarin stimuli in participants with no knowledge of any tone language, and the training study by Li and DeKeyser (2017) found that musical tonal ability correlated with performance in both production and perception.

At pre-test, *Melody Memory* did *not* predict the performance of either group in either tasks, with evidence for the null, while *Beat perception* was found to be a strong predictor, although for Pitch Contour Perception Test this was only found in the MLP group (with evidence for the null for the NP group). It is unclear why there was no effect of *Melody Memory* on either the group. This task captures both the short-term memory and the pitch discrimination aspect of individuals. Participants needed to hold in memory the first piece of music, and compare with the second one played later. The only cue they can rely on is the pitch difference between the two pieces of music. Given that there is a short term memory component, the lack of effect of MM partially corresponds to the results from Working Memory discussed in section 4.5.2.2.2, where there was no evidence for a predictive relationship for three out four Working Memory measures. However this task is similar to one of the music tasks used by Li and DeKeyser (2014), so it is unclear why they found this task to be predictive and the current study didn't. As for *Beat perception*, this task captures participants' sensitivity to rhythm patterns. Recall Bhatara et al. (2015) has reported a correlation between rhythmic perception and longer foreign experience. It could be the case that participants who do better in discriminating beat patterns can better discriminate subtle duration differences of the tones. However for naïve

participants, this ability may be less useful in the Pitch Contour Perception Test at pre-test as without prior knowledge of Mandarin, they can't know which tone should be longer. In contrast, in the Four Interval Oddity task they hear both target and distractor tones so that tone duration could potentially be a cue they use to discriminate the target from the distractor.

Turning to pre-to-post improvement, *Melody Memory* seemed to be important in Four Interval Oddity task while *Beat perception* was a reliable predictor in the Pitch Contour Perception Test. In both cases this was true only for the NP group, a point to which I return below. In terms of the different relationship between the *ID measures* and the outcome measures, the reason may be attributed the nature of these tasks. As suggested above, *Melody Memory* mainly assesses the ability to hold pitch patterns in the mind and discriminate between pitch patterns; this is perhaps most similar to the Four Interval Oddity task, during which participants also need to temporarily hold the tone information in mind for later comparison. Thus it may be natural that a relationship was found between these two measures. The relationship between *Beat perception* and the Pitch Contour Perception Test may indicate that naïve participants who are better at detecting length distinctions between stimuli improve more in their ability to identify the tones. As suggested above, this may indicate length cues indeed play a role in learning the different tones during training. Turning to the group difference, the lack of a predictive role for the MLP group with evidence for the null in some cases (*Melody Memory* predicting improvement in the Four Interval Oddity task) suggest that musical ability may be most relevant in the early stages of learning Mandarin, with attention-related abilities more helpful in later stages. This will be discussed in more detail in the next section.

#### 4.5.2.3 *Summary of findings for ID tasks across measures*

In the current study, different patterns across different *ID measures* were found. Pre-test score from the *Pitch Contour Perception Test* seems to be useful as a predictor for participants already learning Mandarin, both predicting their baseline performance *and* predicting how much can benefit from training. For working memory, there was evidence that *Digit Span Backward* predicted the NP group's Pinyin production at pre-test, however this likely results from the fact that this is essentially a non-word reading task for this group. More critically, *Digit Span Backward* also predicted the improvement of tone production while *Digit Span Forward* predicted improvements in Four Interval Oddity task for the NP group. These two findings could reflect a role for WM- in particular the phonological loop - in tone learning for naïve participants, although it is not clear why it is *Digit Span Backward* in production task and *Digit Span Forward* in the perception task, or why *Digit Span Forward* was not predictive in the Pitch Contour Perception Test. *Letter Number Sequencing* predicted the pre-test performance of the MLP group. This is most likely due to that this task also taps attention switching abilities. The most intriguing finding of the current study was that for those already studying Mandarin, a range of attention measures are very strongly predictive of both of baseline performance *and* the ability to benefit from training materials. However, this didn't apply to naïve participants. Musical ability was found to play a role in predicting baseline performance in the perception tasks to some extent for both participant groups. However in terms of ability to benefit from training, it only predicted the performances of the NP group.

Combining these findings together, the current study seemed to indicate that for naïve participants who just started learning Mandarin, their ability to benefit from phonetic training is mostly predicted by Musical Ability, and perhaps also to some extent Working Memory, whereas for those already learning Mandarin it is Attention that is important. When English speaking participants try to process Mandarin, a tonal language, they'll perceive both the tonal and the phonetic information of the Mandarin words. For naïve participants, Musical Ability

may be important in allowing them to distinguish the tones and thus benefit from training. In contrast, Mandarin learners already have acquired some knowledge of Mandarin tone pitch patterns and thus have less difficulty identifying tones. It is possible that the MLP group are focusing on learning the new Mandarin words, including both the segmental phonology as well as the tones, which require them to allocate attention to different aspects of the stimuli. In contrast, the NP group may be primarily focusing on the new aspect of the stimuli which is the tones, and Musical Ability is the most relevant for this process.

#### *4.5.3 Limitations and future directions*

There are several potential limitations of the current study. Firstly, although the current study provides some preliminary results regarding the difference between individuals who have studied Mandarin compared to those who are naïve to the language, interpretation is limited by the fact that, due to the recruitment method, there are differences between the MLP and the NP groups beyond the fact that one group has studied Mandarin. Specifically, the MLP group are younger. Thus although it was found that the MLP group outperformed the NP group on measures of Working Memory, Attention and Musical ability it is unclear whether differences between the groups are due to their experience learning the language versus other differences between the samples. An attempt was made to explore the possible role of age, and analyses did suggest that this did not play a role in the current experimental tasks (with evidence for the null). It may be that the age difference in the current study isn't large enough to be important: Studies looking at individual differences have generally reported a cognitive advantage of younger adults (around 25 years old) over middle-aged groups (around 50 years old) such as working memory (Wild-Wall, Falkenstein & Gajewski, 2011) and selective attention (Barr & Giambra, 1990) but there hasn't been any report on cognitive differences between 20-year-olds and 30-year-olds. Being current students in the environment of university might also make the

MLP group more accustomed to undertake experiments and to be more motivated (for a discussion on potential theoretical problems caused by using student sample, see Henry (2008)). Thus any replication of the current study should use a matched sample from the same age group and with the same level of education (e.g. students from the same university in the same year). This would help to confirm that the differences found in the role of cognitive measures did not result from age differences.

Secondly, some of the performance measures used in the current design might not be well suited to capture the learning of the participants. In particular, the production measures were too easy for the MLP group, leading to ceiling effects and potentially obscuring relationships with *ID measures*. In a follow-up study, acoustic measures could be used. For example, studies have shown that fundamental frequency is different in English and Mandarin, with F0 used in single utterances having higher maximums and means, and larger ranges for Mandarin speakers (Keating & Kuo, 2012). Gottfried and Ouyang (2005) compared the F0 of tone 4 produced by native English speakers and the one produced by native Mandarin speakers as a measure of how well English speakers can imitate Mandarin tones. They found that musicians performed better than general individuals. The current study might have found a link between musical ability and production if it had used this measure of F0 similarity. Although the NP group did demonstrate improvement in the production measures, as discussed in Section 4.5.2, the choice of using a reading task with largely familiar orthography may also limit the current results. An alternative choice would be using the training paradigm based on learning word-picture mappings and use production tasks such as Word Repetition and Picture Naming as in Study 1 and 2.

Thirdly, although the current perception measures led to a better range of results for both participant groups, it is possible that more sensitive tasks using RT measures could have

detected further relationships with individual differences, especially for the MLP group. For example Ou et al. (2015) and Ou and Law (2017) reported that attention and working memory predicted participants' processing speed of Cantonese tones. In the current study, although a strong pattern between Attention measures and the MLP group has been revealed, no such relationship was found with working memory. Perhaps including RT for Four Interval Oddity task and PCPT might allow us to see more clearly how Working Memory affects Mandarin tone learning (Although using RT would be less helpful for the NP group considering their overall accuracy was much lower to begin with making RT inappropriate). Using RT could also potentially shed further light on the pattern of results found in Attention measures. If the explanation given above is correct, and consistent with Li and Francis (2014) and Zou et al. (2017), that the current MLP group did attend more to both tones and phonology segments, then the expected pattern is to see quicker response times from the NP group over the MLP group, despite the greater accuracy of the latter group.

When drawing conclusions from the current findings, it is important to acknowledge that this work is largely exploratory and as such involved a large number of measures and tests. A potential concern is the lack of control for the testing of multiple hypotheses which makes the *p*-values hard to interpret. Although Bayes Factors were the key method of interest and theoretically they remain a valid measure of evidence regardless of the number of hypotheses tested. Nevertheless, Type 1 error could be a potential problem. In the analysis I also attempted to create and use more robust measures, i.e. the composite score for each type of *ID measures* which has been used in previous literature (e.g. Ou et al., 2015). However, it was unclear how useful these composite scores were. The general pattern was that wherever at least one of the subtests was predictive, this was also reflected in the composite score (with the exception of Music composite score, where it always showed the same pattern as the *Beat perception* score).

An alternative method to use with multiple measures would be Principal Component Analysis (PCA, e.g. Li and DeKeyser, 2017; Meda et al., 2009), which allows the grouping of factors which capture similar variance. I did initially attempt this for the current analyses (see script [https://osf.io/j6s7w/?view\\_only=497e0e8ee7ff4e7387984690eafd4b5a](https://osf.io/j6s7w/?view_only=497e0e8ee7ff4e7387984690eafd4b5a)), however the patterns were hard to interpret. Despite this, in the results as presented, there are many places where the pattern of results is extremely clear and strong- for example the very large amount of evidence that Attention predicts perceptual tone learning for the MLP group, which held across a range of measures. However, to corroborate the findings it would be necessary to replicate the study using with the same set of tests (or at least the subset of tests which were shown to be predictive). An advantage of this replication would be that the effect sizes found in the current data could be used to inform H1 in the Bayes Factor analyses. These analyses could be pre-registered prior to collecting the data (van't Veer & Giner-Sorolla, 2016).

Statistical power is another potential concern in the current study. Only 60 participants were recruited, due to the time consuming nature of the training component to the study. Ou et al. (2015) also looked at individual difference and tone discrimination ability and they recruited 136 participants, more than twice of the sample size used in the current design. On the other hand, however, some analyses reported above did demonstrate extremely large Bayes Factors (e.g. Attention measures predicted the pre-test performance of the MLP group in PCPT, with Bayes factors greater than 9999, see Table 24). Recall that Bayes Factors can be interpreted continuously and these large numbers indicate that there is great deal of evidence for the effect in question. There were also cases where there was sufficient sample to provide evidence for the null. However where the power was particularly weak was for the comparison between groups where the evidence was very often ambiguous. In particular, in all but two instances, this was always the case wherever there was no overall effect of the *ID measure*. Similarly to

Study 2 (Section 3.3.7), I conducted an additional analysis to estimate the sample size which would be needed to find substantial evidence for H1 for some of the effects (again based on the assumption that the error term would reduce in proportion to  $\sqrt{SE}$ ). The results suggested a sample size approximately seven times bigger than the current one would be necessary if aiming to see a difference between groups. For example, when looking at group difference at pre-to-post improvement in PCPT, 432 participants will be required in order to see an effect.

Finally, future work could explore whether more explicit learning conditions lead to stronger relationship with individual differences. Dating back to Reber, Walkenfeld and Hernstadt (1991), researchers have suggested that learning under implicit circumstances is insensitive to measures of cognitive abilities, such as working memory. They exposed participants to artificial languages and then tested them using grammatical judgement tasks. Those trained under explicit conditions - i.e. they were told that they should focus on grammatical aspects of the stimuli- showed better performance than those exposed under implicit conditions. In addition, a correlation between participants' performance and IQ score was reported only for the first group. Although this study used a measure of IQ, which is different from the types of *ID measures* employed in this study, the results were supported by several follow up studies including some which looked at working memory. For example, Robinson (1997) exposed participants to either an incidental, meaning-focused learning condition or an instructed, form-focused condition with artificial languages. He also measured their auditory working memory capacity. The results suggested that the memory measure was correlated with learning outcomes in the explicit condition only. Erlam (2005) studied New Zealand secondary school students learning object pronouns in L2 French. They created three groups: rule-instruction + practice, practice only and rule-instruction only. Results suggested that instruction benefited language learning. The results also suggested that participants'

working memory measure only correlated with performance for those who received explicit instructions. In the current study, although some basic instructions about Mandarin tones were given before testing, it is hard to define whether this counts as explicit learning, as the experimenter didn't tell participants the training purpose nor the details of Mandarin. According to previous literature (e.g. Tagarelli, Mota & Rebuschat, 2011), "explicit" conditions make it clear to the participants that they are aiming to learn certain rules. They can be presented with the rules directly, or have their attention deliberately drawn to relevant aspects of the stimuli. In contrast, implicit conditions do not inform participant the true purpose of the experiment, so that participants do not know that they need to specifically learn something, or that they will be tested. The current design appears to be a mix of implicit/explicit conditions, however it is possible that if they were more directly instructed to focus on the tones this might lead to more explicit learning, and thus their learning of tones might show clearer correlation with individual difference measures.

#### 4.5.4 Conclusion

In conclusion, the current study found relationships between Mandarin tone learning and individual differences. In general, the results had limited findings for production, due to ceiling effects in the MLP group, however there was some evidence for the NP group for a role of one of the Working Memory measures: *Digit Span Backward* predicted post-training improvement in tone and baseline performance in Pinyin. For perception, it was found that for naïve learners who just started learning Mandarin, Musical Ability was the dominant predictor of their improvement from pre- to post- test, although there was some evidence that one of the Working Memory measure, *Digit Span Forward* also played a role. For experienced learners who have already learned Mandarin for 18 months in university, there is a strong correlation with various measures of Attention (and a working memory measures which is thought to

reflect attention: *Letter Number Sequencing*). Attention predicted both their pre-test performance and their pre to post improvement. The pre-test score from the *Pitch Contour Perception Test*, which has been used as individual difference measure in previous work, found to be predictive both of pre-test performance, *and* for pre-to-post improvement, in the Four Interval Oddity task, but only for participants who were already learning the language.

#### **4.6 Principal component analysis**

Analyses in Section 4.4 revealed a complex set of relationships across different individual difference measures. With the goal of finding more general patterns, I also used composite scores, combining Working memory, Attention and Music measures. However, as discussed in Section 4.5 this was not very informative as the effects of each composite scores was generally aligned with the contained individual difference measures which had the strongest effect for that group in that task (e.g. the effect of Working Memory composite score is the same as the effect of *Letter Number Sequencing* when it predicts the performance of Mandarin Learner Participants in the Four Interval Oddity task). Thus, I also attempted a principal component analysis as a way of supplementing the previous analysis. The goal is to determine which of the measures can be grouped together on a statistical basis.<sup>16</sup> Principal component analysis (PCA) analyses a data matrix representing observations described by several dependent variables which are thought to be inter-correlated. Its goal is to extract the important information from the data matrix and to express this information as a set of new orthogonal variables called principal components (Abdi & Williams, 2010). These principal components are obtained as linear combinations of the original variables. The first principal component is required to have the largest possible variance (i.e., therefore this component will

---

<sup>16</sup> Many thanks to examiners Paul Iverson and Ocke-Schwen Bohn who suggested this additional analyses and offered advice regarding PCA.

“explain” or “extract” the largest part of the data matrix). The second component is computed under the constraint of being orthogonal to the first component and to have the largest possible variance. The other components are computed likewise (Wold, Esbensen & Geladi, 1987). The values of these new variables for the observations are called principle component scores, these scores can be interpreted as the combination of the contribution from different DVs and the size of the contribution is reflected in the loadings of each DV. The current analysis used PCA to transform the 11 individual difference measures across Working Memory, Attention and Musical Ability to a set of four principal components. Importantly, data from the experiment performances measures (i.e. from Training, Four Interval Oddity, Pitch Contour Perception Test and Pinyin Naming) were not included in these analyses that identifies the principal components. (It should be noted that although pre-test scores from the Pitch Contour Perception Test were used as an individual difference measure in the previous analyses, it was not included here since then the components could not be used as predictors in the analysis of that data, and in any case it is of a rather different nature from the other cognitive individual difference measures). Separate PCA analyses were run for the NP group and the MLP group, given that Section 4.4.2 revealed that the participants groups differed substantially on these measures so the relevant components may therefore be different for the two groups. In order to keep the measures on the same scale, z scores were computed for each ID measures (and for *Visual Elevator*, *Telephone Search* and *Telephone Search while Counting*, where smaller RT measures represented better performance, the sign was changed before computing the Z score) before they were entered into the analyses. The analysis reported in this section used the “principal” function in the “psych” package in R (Revelle, 2019).

Bartlett’s test of sphericity (naïve participants:  $X^2(55) = 157.60, p < .001$ ; Mandarin learner participants:  $X^2(55) = 219.37, p < .001$ ) indicated that the correlation structure was

adequate for principle component analyses. The Kaiser's criterion of eigenvalues greater than 1 (see Field, 2009) yielded a four-factor solution as the best fit for the data, accounting for 81.89% of the variance for the NP group and 75.82% of the variance for the MLP group. The varimax rotation was applied. The results of this analysis are presented in Table 26 (for the NP group) and Table 27 (for the MLP group). It can be seen that the principal components formed are different for the NP and the MLP group.

For the NP group: (a) the first component demonstrated a clear loading of memory-related tasks i.e. *Arithmetic, Melody Memory, Digit Span Forward & Digit Span Backward*, bearing in mind that the Melody Memory task taps working memory as well as and musical ability. This component explained 30% of the total variance and I name it the Working Memory (WM) Component (b) the second component loaded on three attention measures. It should be noted that all these three measures are RT measures (*Telephone Search, Visual Elevator, Telephone Search while Counting*). Recall that I used Z scores (and changed the sign) so that the scale of the RT measures should not be what underpins this grouping. This component explained 25% of the total variance and I name it the Attention Reaction Time (ART) Component. The third component included a combination of working memory measures (*Letter Number Sequencing, Digit Span Backward*) and one attention measure (*Elevator Counting with Distraction*). This component explained 16% of the total variance and I name it the Working Memory + Attention (WM+A) Component (d) the fourth component included one attention measure (*Elevator Counting with Reversal*) and one musical ability measure (*Beat Perception*). This is relatively harder to interpret as the musical ability component showed a negative loading. This component explained 11% of the total variance and I name it the ECR-BP Component.

For the MLP group: (a) the first component demonstrated a clear loading attention measures (*Elevator Counting with Reversal, Telephone Search while Counting & Telephone Search*). This component explained 35% of the total variance and I name it the Attention Component (b) the second component loaded on one working memory measure (*Letter Number Sequencing*), one musical ability measure (*Beat Perception*) and one attention measure (*Visual Elevator*). Recall from previous discussion that *Letter Number Sequencing* is believed to have a large attentional component aspect. This component explained 20% of the total variance and I name it Attention + Music Ability (A+MA) Component (c) the third component included a combination of working memory measures (*Digit Span Forward, Digit Span Backward*) and one attention measure (*Elevator Counting with Distraction*). Interestingly, this component resembled the WM+A Component found for the NP group (*Letter Number Sequencing, Digit Span Backward & Elevator Counting with Distraction*). This component explained 12% of the total variance and I name it the Working Memory + Attention (WM+A) Component (d) the fourth component included one Working Memory measure (*Arithmetic*) and one musical ability measure (*Melody Memory*). This is relatively harder to interpret. Although both tasks reflect aspects of Working Memory, the *Arithmetic* task showed a negative loading. This component explained 9% of the total variance. I name it the MM-ARI Component.

After extracting the components, I then carried out a new series of analyses using logistic mixed effects models similar to those reported in Section 4.4 but with the extracted component, instead of individual difference measures and composite scores, as the predictors. For these analyses, unlike in the previous analysis, I do not use separate models for each predictor but instead, enter all of the components into the same model as factors (noting that they are orthogonal). I also include the effect of *test-session/training-session* and interaction between each component as well as the main effect of *test-session/training-session*. In contrast

to the previous analyses, separate models were created for the MLP and the NP group since the components, i.e. the predictors, are different for the two groups. As a result, this new set of analyses has 10 models (two for each of the performance tasks: Training, Pinyin Reading Tone accuracy, Pinyin Reading Pinyin accuracy, Four Interval Oddity & Pitch Contour Perception Test).

For these analyses I did not compute Bayes factors as it is unclear at this stage how to inform the H1's for the components. Instead, I use the frequentist  $p$ -values provided as the measure of inference. For each of the models, as there are four factors included (component 1-4), I apply the Bonferroni correction for four tests i.e.  $p$  value at 0.05 would now be 0.0125 and  $p$  value at 0.01 would now be 0.0025.

*Table 26* Principle Component Analysis of the individual differences measures of Working Memory, Attention and Musical Ability for the naïve participants group. Loadings larger than 0.40 are in bold.

Individual differences Measures	Components			
	WM	ART	WM+A	ECR-BP
Digit Span Forward	<b>0.95</b>	-0.01	-0.01	0.13
Digit Span Backward	<b>0.49</b>	-0.39	<b>0.69</b>	0.01
Letter Number Sequencing	-0.09	0.12	<b>0.86</b>	0.04
Arithmetic	<b>0.97</b>	0.14	0.04	-0.06
Elevator Counting with Distraction	0.28	0.31	<b>0.53</b>	0.38
Visual Elevator	-0.10	<b>0.86</b>	0.17	0.24
Elevator Counting with Reversal	-0.05	0.30	-0.05	<b>0.82</b>
Telephone Search	0.06	<b>0.93</b>	0.21	-0.08
Telephone Search while Counting	0.12	<b>0.84</b>	-0.27	-0.01
Beat Perception	-0.01	0.19	-0.15	<b>-0.75</b>
Melody Memory	<b>0.96</b>	-0.02	0.13	-0.08

Individual differences Measures	Components			
	WM	ART	WM+A	ECR-BP
Variance explained	30.2%	25.4%	15.7%	10.6%
Eigen value	3.33	2.79	1.73	1.16

Table 27 Principle Component Analysis of the individual differences measures of Working Memory, Attention and Musical Ability for the Mandarin learner participants group. Loadings larger than 0.40 are in bold.

Individual differences Measures	Components			
	Attention	A+MA	WM+A	MM-ARI
Digit Span Forward	0.00	0.15	<b>0.87</b>	0.14
Digit Span Backward	-0.32	0.28	<b>0.81</b>	-0.09
Letter Number Sequencing	0.07	<b>0.88</b>	0.18	-0.16
Arithmetic	-0.12	-0.06	0.09	<b>-0.79</b>
Elevator Counting with Distraction	0.39	-0.06	<b>0.77</b>	0.11
Visual Elevator	0.33	<b>0.74</b>	0.11	0.15
Elevator Counting with Reversal	<b>0.85</b>	0.37	-0.07	-0.05
Telephone Search	<b>0.71</b>	0.32	0.28	0.25
Telephone Search while Counting	<b>0.85</b>	0.15	-0.07	-0.05
Beat Perception	0.35	<b>0.79</b>	0.05	0.09
Melody Memory	-0.09	-0.03	0.26	<b>0.75</b>
Variance explained	34.5%	19.8%	12.5%	9.1%
Eigen value	3.80	2.18	1.37	1.00

#### 4.6.1 Training

Both models converged (in contrast to the analyses with separate measures as predictors reported in 4.4.3 above). The results are summarised in Table 28. It can be seen that there was no effect found for any component for the naïve participants. For the Mandarin learner participants, it can be seen that the 1st component, Attention Component predicted both Mandarin learner participants' performance in the first training session and their learning progress across sessions. The third component, WM+A also predicted their learning slope.

Table 28 Regression analysis for Training task with principle components as predictors. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis).

Components	$\beta$	<i>SE</i>	<i>Z</i>	<i>p</i>
Naïve participants				
Working Memory Component	-0.04	0.26	-0.16	0.870
Working Memory Component by Session	0.08	0.08	1.02	0.307
Attention RT Component	0.12	0.27	0.45	0.654
Attention RT Component by Session	-0.05	0.08	-0.65	0.52
WM+A Component	0.21	0.27	0.79	0.429
WM+A Component by Session	-0.07	0.08	-0.90	0.367
ECR-BP Component	0.31	0.26	1.16	0.246
ECR-BP Component by Session	-0.15	0.08	-2.04	0.041
Mandarin Learner participants				
<b>Attention Component</b>	<b>0.69</b>	<b>0.20</b>	<b>3.36</b>	<b>&lt;0.002</b>
<b>Attention Component by Session</b>	<b>0.19</b>	<b>0.04</b>	<b>4.45</b>	<b>&lt;0.002</b>
A+MA Component	0.22	0.20	1.10	0.271
A+MA Component by Session	0.07	0.04	1.55	0.120
WM+A Component	-0.13	0.20	-0.62	0.534
<b>WM+A Component by Session</b>	<b>0.12</b>	<b>0.04</b>	<b>2.98</b>	<b>0.003</b>
MM-ARI Component	-0.21	0.20	-1.04	0.298
MM-ARI Component by Session	0.04	0.05	0.79	0.431

#### 4.6.2 Pinyin Reading

##### 4.6.2.1 Tone accuracy

The results are summarised in Table 29. It can be seen that the third component, WM+A predicted naïve participants' improvement in Pinyin Reading, tone accuracy. For the Mandarin learner participants, no effect was found for any component.

Table 29 Regression analysis of Pinyin Reading, tone accuracy with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis).

Components	$\beta$	$SE$	$Z$	$p$
Naïve participants				
Working Memory Component	-0.06	0.06	-1.00	0.317
Working Memory Component by Session	0.35	0.13	2.63	0.008
Attention RT Component	-0.003	0.06	-0.06	0.955
Attention RT Component by Session	-0.20	0.13	-1.51	0.132
WM+A Component	-0.04	0.06	-0.76	0.447
<b>WM+A Component by Session</b>	<b>0.58</b>	<b>0.13</b>	<b>4.45</b>	<b>&lt;0.002</b>
ECR-BP Component	0.01	0.06	0.12	0.908
ECR-BP Component by Session	0.01	0.13	0.08	0.937
Mandarin Learner participants				
Attention Component	0.01	0.07	0.08	0.937
Attention Component by Session	0.00	0.10	-0.03	0.973
A+MA Component	0.02	0.07	0.34	0.735
A+MA Component by Session	0.18	0.11	1.63	0.103
WM+A Component	-0.01	0.07	-0.12	0.905
WM+A Component by Session	-0.09	0.11	-0.76	0.449
MM-ARI Component	0.04	0.07	0.65	0.519
MM-ARI Component by Session	-0.13	0.11	-1.16	0.247

#### 4.6.2.2 Pinyin accuracy

The results are summarised in Table 30. It can be seen that WM+A predicted naïve participants' baseline performance in Pinyin Reading, tone accuracy. For the Mandarin learner participants, no effect was found for any component.

Table 30 Regression analysis of Pinyin Reading, pinyin accuracy with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis).

Components	$\beta$	SE	Z	p
<b>Naïve participants</b>				
Working Memory Component	0.15	0.06	2.35	0.019
Working Memory Component by Session	0.02	0.11	0.15	0.879
Attention RT Component	-0.03	0.06	-0.53	0.596
Attention RT Component by Session	0.15	0.10	1.45	0.147
<b>WM+A Component</b>	<b>0.28</b>	<b>0.06</b>	<b>4.51</b>	<b>&lt;0.003</b>
WM+A Component by Session	-0.23	0.10	-2.18	0.029
ECR-BP Component	-0.02	0.06	-0.33	0.744
ECR-BP Component by Session	0.07	0.10	0.64	0.521
<b>Mandarin Learner participants</b>				
Attention Component	0.01	0.07	0.19	0.850
Attention Component by Session	0.05	0.13	0.37	0.710
A+MA Component	-0.05	0.07	-0.66	0.510
A+MA Component by Session	0.14	0.14	1.01	0.311
WM+A Component	0.04	0.07	0.63	0.527
WM+A Component by Session	-0.19	0.14	-1.37	0.171
MM-ARI Component	0.004	0.07	0.06	0.953
MM-ARI Component by Session	-0.08	0.14	-0.55	0.584

#### 4.6.3 Four Interval Oddity task

The results are summarised in Table 31. For naïve participants, WM Component predicted their pre to post improvement. For the Mandarin learner participants, the Attention Component predicted both their performance at pre-test and improvement after training. A+MA Component predicted only their pre-test performance.

Table 31 Regression analysis of Four Interval Oddity with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis).

Components	$\beta$	$SE$	$Z$	$p$
Naïve participants				
Working Memory Component	0.18	0.08	2.17	0.030
<b>Working Memory Component by Session</b>	<b>0.37</b>	<b>0.10</b>	<b>3.61</b>	<b>&lt;0.002</b>
Attention RT Component	0.18	0.08	2.22	0.026
Attention RT Component by Session	-0.10	0.10	-1.02	0.306
WM+A Component	-0.07	0.08	-0.90	0.368
WM+A Component by Session	0.18	0.10	1.84	0.066
ECR-BP Component	-0.04	0.08	-0.54	0.592
ECR-BP Component by Session	0.03	0.10	0.30	0.763
Mandarin Learner participants				
<b>Attention Component</b>	<b>0.30</b>	<b>0.07</b>	<b>4.28</b>	<b>&lt;0.002</b>
<b>Attention Component by Session</b>	<b>0.36</b>	<b>0.08</b>	<b>4.44</b>	<b>&lt;0.002</b>
<b>A+MA Component</b>	<b>0.51</b>	<b>0.08</b>	<b>6.50</b>	<b>&lt;0.002</b>
A+MA Component by Session	0.01	0.10	0.10	0.923
WM+A Component	0.09	0.07	1.20	0.231
WM+A Component by Session	0.11	0.08	1.40	0.160
MM-ARI Component	-0.18	0.08	-2.25	0.024

MM-ARI Component by Session	-0.003	0.11	-0.03	0.977
-----------------------------	--------	------	-------	-------

#### 4.6.4 Pitch Contour Perception Test

The results are summarised in Table 32. For naïve participants, *ECR-BP* component predicted their learning progress. For the Mandarin learner participants, it can be seen that both *Attention* Component and *A+MA* Component predicted their performance at pre-test.

Table 32 Regression analysis of Pitch Contour Perception Test with principle components for the naïve participants and the Mandarin learner participants. Significant results are marked in bold. These have  $p < .0125$  (at  $\alpha = 0.05$  corrected for four comparisons per hypothesis).

Components	$\beta$	<i>SE</i>	<i>Z</i>	<i>P</i>
Naïve participants				
Working Memory Component	0.09	0.09	1.02	0.307
Working Memory Component by Session	-0.06	0.08	-0.74	0.462
Attention RT Component	-0.14	0.09	-1.62	0.106
Attention RT Component by Session	0.08	0.08	0.95	0.343
WM+A Component	-0.01	0.09	-0.14	0.885
WM+A Component by Session	0.001	0.08	0.01	0.991
ECR-BP Component	0.13	0.09	1.50	0.133
<b>ECR-BP Component by Session</b>	<b>-0.38</b>	<b>0.08</b>	<b>-4.54</b>	<b>&lt;0.002</b>
Mandarin Learner participants				
<b>Attention Component</b>	<b>0.36</b>	<b>0.05</b>	<b>6.91</b>	<b>&lt;0.002</b>
Attention Component by Session	0.14	0.08	1.80	0.072
<b>A+MA Component</b>	<b>0.34</b>	<b>0.06</b>	<b>6.13</b>	<b>&lt;0.002</b>
A+MA Component by Session	0.12	0.08	1.37	0.170
WM+A Component	0.10	0.05	1.99	0.046
WM+A Component by Session	0.07	0.07	0.99	0.320

MM-ARI Component	-0.07	0.06	-1.28	0.201
MM-ARI Component by Session	0.09	0.08	1.06	0.288

---

#### 4.6.5 Discussion

Table 33 (naive participants) and Table 34 (Mandarin learner participants) summarise the results of the analysis in Section 4.4 and the new analyses. The key difference is that in the original analyses, separate individual difference measures were used as predictors (with each one entered into a separate model predicting relevant performance measures) whereas in the new analyses above, the predictors used were four components identified using PCA (which, because they are orthogonal, could all be entered into the same model for each task). Despite the fact that I also used different inference criteria in these two sets of analyses - i.e. Bayes factors for the first set of analyses (without corrections for multiple hypotheses) - and frequentists  $p$ -values (with correction for multiple hypotheses) for the second analyses, it can be seen that these two sets of analyse broadly match with each other. In fact, wherever I found evidence for a measure being predictive in the original analysis, the component most heavily weighted on that factor is predictive in the second analysis. There are only two exceptions: firstly, the original analyses found that *Elevator Counting with Distraction* was a good predictor for both the NP and the MLP groups in Four Interval Oddity task. However, in the current components, this particular measure most strongly contributed to the WM+A Component for both the NP and the MLP groups and this component was not predictive in the current analyses. However, it should be noted that this component was mainly loaded on *Digit Span Backward* (0.69) and *Letter Number Sequencing* (0.86) for the Naïve participants, and *Digit Span Forward* (0.87) and *Digit Span Backward* (0.81) for the Mandarin Learner participants. The potential effect of *Elevator Counting with Distraction* may be overcome by

other loadings. This demonstrates that if we rely on analyses using principal components as predictors, we may potentially lose our ability to see the effects of a particular individual measure, if that measure does not end up making a large contribution to any of the top components identified. Secondly, in the original analyses, *Elevator Counting with Reversal* and *Telephone Search* were predictive in Pitch Contour Perception Test for the MLP group. However in the new analyses the relevant Attention Component (85% from *Elevator Counting with Reversal*, 71% from *Telephone Search*, 85% from *Telephone Search while Counting*) was not predictive. In this case, the difference seems to be due to the different inference criteria used. In the previous analyses these two individual difference measures had relatively smaller Bayes Factors compared to some of our other measures (*Elevator Counting with Reversal*: 6; *Telephone Search*: 30), while the p-value for Attention Component by session interaction was 0.07, which is close to significance without correction.

*Table 33* Summarised results from previous analyses using individual measures as predictors in section 4.4 and current analyses using principal components as predictors (section 4.6.1-4.6.4) for naïve participants. Where cells are grey represents no predictors were identified.

Task	Session	Naïve participants	
		Predictors identified in previous analysis	Predictors identified in current analysis
Pinyin Reading – Tone accuracy	Pre-test		
	Pre-post Improvement	Digit Span Backward	Component 3: WM+A [Loadings from: Digit Span Backward (0.69), Letter Number Sequencing(0.86), Elevator Counting with Distraction, (0.53)]
Pinyin Reading – Pinyin accuracy	Pre-test	Digit Span Backward	Component 3: WM+A [Loadings from: Digit Span Backward (0.69),

			Letter Number Sequencing(0.86), Elevator Counting with Distraction, (0.53)]
	<b>Pre-post Improvement</b>		
<b>Four Interval Oddity</b>	<b>Pre-test</b>		
	<b>Pre-post Improvement</b>	Digit Span Forward Elevator Counting with Distraction Melody Memory	Component 1: WM [Loadings from: Digit Span Forward (0.95), Digit Span Backward (0.49), Arithmetic (0.97), Melody Memory (0.96)]
<b>Pitch Contour Perception Test</b>	<b>Pre-test</b>		
	<b>Pre-post Improvement</b>	Beat Perception	Component 4: ECR-BP (negative $\beta$ ) [Loadings from: Elevator Counting with Reversal (0.82), Beat Perception (-0.75)]
<b>Training</b>	<b>First session</b>	No analysis was performed due to model convergence issues	
	<b>Improvement through Training</b>		

Table 34 Summarised results from previous analyses using individual measures as predictors in section 4.4 and current analyses using principal components as predictors (section 4.6.1-4.6.4) for Mandarin learner participants. Where cells are grey represents no predictors were identified.

Task	Session	Mandarin Learner	
		Predictors Identified in Bayesian analysis	Components Identified in Principal Component analysis
<b>Pinyin Reading – Tone accuracy</b>	<b>Pre-test</b>		
	<b>Pre-post Improvement</b>		
<b>Pinyin Reading – Pinyin accuracy</b>	<b>Pre-test</b>		
	<b>Pre-post Improvement</b>		
<b>Four Interval Oddity</b>	<b>Pre-test</b>	Letter Number Sequencing; Elevator Counting with Reversal; Visual Elevator; Telephone Search; Telephone Search while Counting; Beat Perception	Component 1: Attention [Loadings from: Elevator Counting with Reversal (0.85), Telephone Search (0.71), Telephone Search while Counting (0.85)]  Component 2: A+MA [Loadings from: Letter Number Sequencing (0.88), Visual Elevator (0.74), Beat Perception (0.79)]
	<b>Pre-post Improvement</b>	Elevator Counting with Reversal; Elevator Counting with Distraction; Telephone Search	Component 1: Attention [Loadings from: Elevator Counting with Reversal (0.85), Telephone Search (0.71), Telephone Search while Counting (0.85)]
<b>Pitch Contour Perception Test</b>	<b>Pre-test</b>	Letter Number Sequencing; Elevator Counting with Reversal; Visual Elevator; Telephone Search; Telephone Search while Counting; Beat Perception	Component 1: Attention [Loadings from: Elevator Counting with Reversal (0.85), Telephone Search (0.71), Telephone Search while Counting (0.85)]  Component 2: A+MA [Loadings from: Letter Number Sequencing (0.88), Visual Elevator (0.74), Beat Perception (0.79)]

	<b>Pre-post Improvement</b>	Elevator Counting with Reversal; Telephone Search	
<b>Training</b>	<b>First session</b>	No analysis was performed due to model convergence issues	Component 1: Attention [Loadings from: Elevator Counting with Reversal (0.85), Telephone Search (0.71), Telephone Search while Counting (0.85)]
	<b>Improvement through Training</b>		Component 1: Attention [Loadings from: Elevator Counting with Reversal (0.85), Telephone Search (0.71), Telephone Search while Counting (0.85)]  Component 2: A+MA [Loadings from: Letter Number Sequencing (0.88), Visual Elevator (0.74), Beat Perception (0.79)]

The next part of the discussion focuses on the contribution of the new analyses: does using the models including factors identified by PCA as predictors provide a clearer picture than the original analyses? Starting with the NP group, there were three components which were predictive: First, the first component, which I named WM Component since it was loaded on measures from the working memory battery -*Digit Span Forward*, *Digit Span Backward*, *Arithmetic* as well as the *Melody Memory*, a musical ability test which also captures aspects of WM capacity - predicted aspects of the production test. Second, pre-to-post improvement in the Four Interval Oddity tasks was predicted by the third component, which I named WM+A, which was also loaded on factors from the working memory battery - *Digit Span Backward* and *Letter Number Sequencing*, and also a task from the attention battery: *Elevator Counting with Distraction*. Do these working memory and attention tasks have anything in common? Both *Digit Span Backward* and *Letter Number Sequencing* involve manipulation of memorised

information (reverse the order of numbers/rearrange the order of numbers and letters) and thus both tasks require the use of attentional resources (although the literature has normally considered only Letter Number Sequencing as a task which taps attention; e.g. Awh et al., 2006). This may explain why they are grouped with *Elevator Counting with Distraction*, which reflects sustained attention. Thus, this component may reflect aspects of working memory which also rely on sustained attention. Four Interval Oddity task requires participants to hold four example words in memory and compare their tones thus this component may be particularly relevant. Finally, the fourth component, which I named ECR-BP, predicted pre- to post- improvement performance in the Pitch Contour Perception test. As noted above, this component is hard to interpret. It was loaded on two tasks which reflect the ability to switch attention between auditory stimuli - *Elevator Counting with Reversal* and *Beat Perception*. However Beat Perception had a *negative* loading while ECR had a positive loading suggesting an inverse relationship between the two. It is unclear why individuals who are *worse* at attention shifting should be better at perceiving musical beats. Moreover caution should be taken in interpreting this pattern, given that the component accounted for only 10.6% of variance and in addition the sample of NP participants was small (N=20). Comparing with the previous Bayes Factor analysis for this test, it is clear that predictive role of this component is dependent on the *Beat Perception* here. The new analysis does not therefore add much to the understanding of this task. In general it remains unclear that the new analyses is very informative. However the fact that two of the components which predict performance comprise mainly WM measures highlights the importance of WM in tone learning paradigm with naïve participants. Finally, it should be noted that none of the components were predictive of training for the NP group (something I was unable to explore in the previous analyses due to convergence difficulties with the model), although in the absence of Bayes Factors we must be cautious in interpreting this null result with this small sample.

For the MLP group, only the first two components played a role in the regression. First, the first component – which I named the Attention Component – was loaded on *Elevator Counting with Reversal*, *Telephone Search* and *Telephone Search while Counting*. These are all tasks from the attention battery which reflect mainly attention switching ability as well as the ability to selectively attend to one stimuli. The second component – which I named the A+MA Component – was loaded on *Letter Number Sequencing*, *Visual Elevator* and *Beat Perception*. Although only Visual Elevator is from the attention battery, these tasks may also reflect attention switching ability: *Letter Number Sequencing* requires participants to switch attention between stored memory chunks (letters/numbers). *Visual Elevator* requires participants to switch between visual stimuli and auditory stimuli, while *Beat Perception* requires participants to switch attention between different auditory stimuli. However, it remains unclear what aspect of cognition differentiates this component from the first component (e.g. why is *Telephone Search while Counting* grouped in this component and not *Visual Elevator*). Similar to the previous Bayes Factor analyses, there are no predictors of production measures. However, the first Attention component is predictive of both baseline performance and pre- to-post improvement in the two perception tests, and also of performance in the first training session and improvement across training sessions. In contrast, the A+MA component only predicted pre-test performance in perception tests (not pre-to-post improvement), and only predicted learning slope (not performance in session one) in Training. However, it should be noted that where we see this component making a contribution may result from the effect of power: there was more data contributing to learning slope (5 sessions worth) than to session one. For the pre- to post test, a more similar amount of data contributes to looking at the effect in session 1 versus the interaction. It is expected that an interaction with *test-session* will be harder to detect than a main effect or simple effect (Leon & Heo, 2009). Thus it seems likely that these differences are due to the A+MA Component, which accounted for less variance in

the PCA analysis, exerting a weaker influence that is harder to detect, rather than there being something fundamentally different about the aspects of cognition which underpin these two components being predictive of baseline performance versus learning and vice versa. More generally, the predictive role of two component loaded on attention is consistent with the previous interpretation that attention is key in predicting performance and learning for Mandarin Learner participants in this tone training paradigm.

The new set of analyses identified some components as important in predicting the performance of the participants in various tasks. However, there are also some other components which were identified by PCA, but which seems to be less relevant for learning Mandarin tones. For the NP group, the second component, which I named the ART Component, was not predictive. This component loads on RT based attention measure (*Visual Elevator, Telephone Search, Telephone Search while Counting*) and thus taps attention processing speed. This may imply that attention processing speed does not play a role in Mandarin tone learning for the NP group. The reason could be that the test measures lack a speed component. However as suggested at the beginning of Section 4.6, unlike with the previous analyses, I did not compute Bayes Factors thus cannot differentiate a true null result- with evidence for H<sub>0</sub>, from type 2 error.

Turning to the MLP group, the third component, which I named WM+A (*Digit Span Forward, Digit Span Backward, Elevator Counting with Distraction*), seems to capture aspects of Working Memory. Although *Elevator Counting with Distraction* was found to be predictive in the first set of analyses, as discussed above, its effect maybe overcome by the other two factors due to having the smallest loading. Interestingly, a similar component *was* predictive for the NP group in the pre-to-post improvement in Pinyin Reading, Pinyin accuracy and pre-test performance in Pinyin Reading, Tone accuracy. The fact that this component was not

predictive for the MLP group may again suggest that as in the previous analyses, Working Memory didn't play a critical role for this group in this task, although without Bayes Factors, I cannot evaluate this null result. The fourth component, MM-ARI (*Arithmetic and Melody Memory*) component, was also not predictive in any task. As noted above it is hard to interpret this component as *Melody Memory* and *Arithmetic* have reversed directions on the weightings. Moreover, it has the Eigen value of exactly 1.00 and only account for 9% of the variance (see Table 27) so it is perhaps not surprising that it is not predictive and in hindsight it may be better to omit this factor from the analysis.

In terms of comparison with previous literature, the majority of the published work using PCA to explore relationships between measures of cognitive differences uses the techniques to find groups of potentially related cognitive factors but does *not* go on to study whether the emerging components are predictive of learning or other aspects of behaviour. For instance, Chan, Lai & Robertson (2006) have also run PCA over participants' scores on the Test of Everyday Attention, however they used the full test battery (whereas the current experiment used a subset) and they did not include any other tests. They found different groupings compared to current study. In the current analysis, *Telephone Search* was grouped with *Telephone Search while Counting* for the MLP group as they both involving attention switching. In Chan et al. (2006), *Telephone Search* was grouped with *Map Search*, as both tasks involving visual selection; while *Telephone Search while Counting* was grouped with *Lottery*, as both tasks involving attending to both visual and auditory stimuli. The fact that different groupings are found demonstrates that the full set of factors entering the model is crucial for what components are formed since by definition PCA will create components based on the contributions from each factor. As for PCA run across batteries of both working memory and attention tests, Machizawa and Driver (2011) explored the relationship between working

memory and attention. They measured visual working memory capacity, visual working memory precision and visual working memory filter ability. Attention was measured using the ANT task, focusing on alerting, orienting and executive aspects. The analysis formed the following three main components, as in the current study they spanned across working memory and attention: visual working memory with alerting scores, visual working memory precision with orienting and visual working memory filtering with executive. It should be noted that the current study actually found the components across working memory, attention and musical ability (e.g. the A+ MA component: *Letter Number Sequencing* (working memory), *Visual Elevator* (attention) and *Beat Perception* (musical ability) found in the MLP group), which was not seen in any previous study. However, although the analysis in Machizawa and Driver (2011) suggested potential relationship between attention and working memory, the researchers did not provide an account of what factors might underpin these components (i.e. explain “why” there are such relationships). To my knowledge, this is the first study which tries to look at grouping across three test batteries (working memory, attention and musical ability), and then consider whether these components can predict learning.

To conclude, I believe including the extra Principle Component analysis does shed a light on the potential relationship between different cognitive abilities, but its use is limited in providing an overall, interpretable pattern of how different cognitive abilities relate to tone perception and tone learning. Using extracted patterns as predictors, this new set of regression models are mainly consistent with the findings from previous analyses that Working Memory is more involved for the NP group and Attention is more involved for the MLP group. However, although components formed generally most related to one of the cognitive categories, the patterns are somewhat hard to interpret as most of the time it remains unclear as why certain tasks are grouped together. Moreover, using extracted components has the cost of losing the

effect of some individual difference measures, either due to the potential thresholding problems (i.e. *Elevator Counting with Distraction*), or because the tasks were grouped with other factors making it hard to interpret its specific effect, as happened with the two Musical Ability measures. This suggests that in future studies, increasing the number of tasks in each cognitive dimension important if PCA is to be used. In particular, having only two Musical Ability tests does not allow for a component to be extracted which accounts for Musical Ability more generally. It would also be useful to apply Bayes Factors to examine the null results. A difficulty here is how to evaluate predicted effect sizes to inform H1 and this should be considered in future work.

## 5. General discussion

The current thesis aims to explore the impact of different factors upon participants' learning of Mandarin lexical tones and words. Chapter 2 reported the first study (Study 1) which had two main goals. Firstly, it explored whether high variability training materials were more effective than low variability training materials when training naïve participants with Mandarin tones. A series of studies (Logan et al., 1991, Lively et al., 1993) have suggested that high variability training is the key in training participants with new phonetic contrasts. These studies have been conducted on learning English contrasts with participants from various linguistic background such as Greek (Giannakopoulou, et al., 2013), German (Iverson, et al., 2008) and Japanese (Bradlow et al., 1999). It is believed that high variability training stimuli are particularly important in allowing participants to generate their learning to new stimuli and new speakers. In Study 1, naïve English speakers learned real Mandarin words accompanied by Mandarin tones in a minimal pair phonetic training paradigm, where they had to use tone to identify the target word. Participants improved their performance throughout training and at test demonstrating generalisation for both new items and new speakers, with this found for both perception and production tasks. However, the study did *not* find any evidence of a high variability advantage for generalisation. In fact, the only difference between the groups was that those trained on low variability stimuli did better during training, and with trained speakers in a test similar to training (Picture Identification). Secondly, Study 1 explored whether individual differences interacted with the ability to benefit from different training conditions. Two previous studies (Perrachione et al., 2011; Sadakata & McQueen, 2014) have reported that when learning Mandarin lexical tones, individual aptitude affected the training efficiency such that high aptitude participants only benefited from high variability stimuli and low aptitude participants only benefited from low variability stimuli. Study 1 therefore also included similar measures of aptitude to those studies. Although in general, positive

relationships were found between one of these aptitude measures and baseline measures of tone production, tone identification and tone discrimination, there was no evidence that the task affected participants' learning from the training materials and, critically, learning was not predicted by the interaction between individual aptitude and their variability condition.

The study reported in Chapter 3 (Study 2) directly followed up the results of Study 1. It is considered that a high variability advantage might not have been seen in the previous study due to the fact that speakers were intermixed during training. Thus a new training condition was introduced which was matched to the previous high variability condition except that rather than intermixing the four speakers all together, each speaker was presented in their own block. This creates a condition with a matched level of speaker variability but much higher trial-by-trial consistency. The results of Study 2 suggested that trial-by-trial consistency does make training easier since the high variability blocked group outperformed the original high variability group. Nevertheless, the low variability group still demonstrated an advantage over this new condition in training and Picture Identification for trained items. However, other than in training, there were no differences between this new blocked condition and the original high variability condition, and, critically, there was still no overall advantage of the high variability stimuli for generalisation and no interaction between the variability condition and individual aptitude measures. This chapter also introduced the use of Bayesian analyses, which were applied to data pooled across the two studies and were used to examine first, whether there was evidence for the null for the benefits of high variability for generalisation; second, whether there was evidence for the null for the interaction between variability and individual aptitude. The results found substantial evidence for the null for the first assumption but not the second, suggesting that it is not possible to rule out the effect of individual aptitude at this stage.

Additional analysis also suggested that a much larger sample would be necessary to find evidence regarding the interaction.

Study 3 further explored the role of individual aptitude in tone learning. The measures of individual difference used in the previous studies were based on those in the previous literature reporting an interaction between aptitude and variability (Perrachione et al., 2011; Sadakata & McQueen, 2014). Nevertheless, only one of these measures (*Pitch Contour Perception Test*) was found to be predictive of participants' ability in baseline performance on tone discrimination and identification, and there was no evidence that it was predictive of participants' of *learning* (i.e. pre-to-post improvement) from any type of training materials. In addition, this measure is a direct assessment of tone identification ability, thus limiting its explanatory value as a measure of individual difference. The new experiment explored a range of individual difference factors that could potentially impact upon the ability to benefit from HVPT on Mandarin tones. Recent studies have suggested that working memory and attention predict Cantonese speakers' perception and production of unfamiliar Cantonese tones (Ou et al., 2015; Ou and Law, 2017) and there is evidence that musical knowledge predicts ability with lexical tones (Li and DeKeyser, 2017), so measures of all of these three cognitive abilities were included, as well as a measure corresponding to the *Pitch Contour Perception Test* used in the previous study. High variability blocked training was used in this study, but no comparison between LV condition was included in order to maximise the sample for this condition and to establish which measures are predictive at least for high variability materials. However, Study 3 did involve two participant types: naïve participants (native English speakers) and Mandarin Learning Participants (second year undergraduates taking a Mandarin programme at SOAS University). A similar phonetic training paradigm was adopted (though with some key differences, in particular the use of Pinyin representations rather than picture

stimuli) and again tests of both perception and production were used. The results suggested an advantage for the group who were already learning mandarin over the naïve participant group on all types of individual difference measures as well as on the performance measures at pre-tests, although there was no clear evidence of an advantage for either group in terms of their ability to improve from pre to post training. In terms of *ID measures* predicting performance, the key finding was that both the *Pitch Contour Perception Test* used in previous studies and the measures of Attention (including a Working Memory measure also measuring attention component) were strongly predictive for the group of participants with previous Mandarin experience, both in terms of baseline results and in the ability to benefit from training. In contrast, Attention did not predict performance for the naïve participants, but there was evidence that their improvement was linked to Musical Ability and possibly Working Memory to some degree. This suggests that there may be a different role for different individual abilities for naïve participants compared with language learners who have already acquired the fundamental rules of the language.

The rest of this discussion will further consider implications of these findings. Firstly, it discusses the implications of not finding a high variability advantage for phonetic training. Secondly, it considers what has been learned about the factors affecting individual aptitude for tone learning at different stages. Thirdly, it considers methodological contribution of the current thesis, specifically the benefit of using Bayes Factors in a phonetic training. Finally, it explores possible directions for future research.

### *5.1.1 The role of speaker variability in phonetic training*

High variability phonetic training, where multiple speakers are used to train non-native speech contrasts, has become standard in the literature of phonetic studies. The benefit of exposure to multiple speakers is highly intuitive: the logic is that if an individual is exposed to

multiple speaker conditions during training, then it should be easier for the individual to adapt to new speakers as the individual has already experienced the variations between speakers. On the other hand, if an individual is only exposed to one speaker during training, dealing with the variability of speakers may be harder. The current thesis further supported that both high variability and low variability training can benefit the learning of Mandarin, by showing that all the groups in Study 1, 2 and 3 demonstrated significant improvement in training and generalise their learning to new speakers and stimuli.

More importantly, the current thesis makes a contribution by showing that – at least for current training materials with real Mandarin words – exposure to multiple speakers is not necessary, or even beneficial, for L2 learners learning lexical tones, with substantial evidence for the null. Comparing this to the literature, it is worth noting that there are actually very few published studies which directly compare LV and HV input. As reported in the Section 1.4, at the time of planning this study, except for the original study (Lively et al., 1993), to my knowledge there were only six studies training non-native phonetic contrasts which involve training participants with both LV and HV materials (Sadakata & McQueen, 2013; Wong, 2012, 2014; Giannakopoulou, et al. (2017) Perrachione et al. (2011) and Sadakata & McQueen, 2014). The first study by Sadakata & McQueen trained native Dutch speakers with Japanese fricative, while the two studies by Wong trained native Cantonese speakers with English /e/ - /æ/ contrasts. All of these studies supported the advantage of HV over LV. However the study by Giannakopoulou, et al. (2017) did *not* find a benefit for HV compared to LV training when training Greek adults and children on English /i:/ - /i/ contrast. As discussed extensively in this thesis, the two studies by Perrachione et al. (2011) and Sadakata and McQueen (2014) on training of lexical tone also did *not* find an overall advantage for high variability training with Mandarin tones. Instead, they found that individual aptitude interacted with the degree of

variability involved in training, such that only individual with high aptitude (i.e. better ability to discriminate tones) can benefit from HV materials. The current study did not replicate this finding, though here Bayesian statistics suggested there isn't enough evidence for the null. Nevertheless, Chapter 3 cautiously discussed whether this different finding could be due to differences in the study designs, for example, greater overall complexity of training materials in the current study. However, it should also be acknowledged that there could be Type 1 error in the previous studies. Since completing these experiments (which have been published as Dong, Brown, Clayards & Wonnacott (2019)) another study has been published which used similar methods and also found null results for both HV benefit and interaction with HV: Zhang, Peng, Li, Minett and Wang (2018) trained native Mandarin speakers on Cantonese tones using either HV (four speaker blocked) or LV training (one speaker only). Six Cantonese tones were paired with 6 Cantonese monosyllabic words. Participants were trained on both perception and production using a word repetition task where the word was played and the tone diacritics were displayed with traditional Mandarin characters. Participants then need to repeat the word and they were explicitly asked to focus on the tone they heard and the diacritic symbol on the screen. The words were played sequentially in a fixed order (e.g. /fən/, tone 55; /fən/, tone 33; /fən/, tone 22;...). Similar to Perrechione et al. (2011) (and the Pitch Contour Perception Test in the current study), they used a Cantonese tone identification task at pre-test, in which participants needed to identify six Cantonese tones (matched to diacritics), as their measure of "aptitude". Their HV group demonstrated similar learning to the LV group, and there was no interaction with their aptitude measure. This corroborates the current results, although they did not employ Bayes Factors or other statistics which could help to evaluate these null results.

### 5.1.2 *Factors affecting individual aptitude for tone learning*

Literature in L2 learning has explored which factors predict successful learning. Factors considered have been affective qualities such as motivation (for a summary, see Gardner, 2014), personality factors such as self-esteem (for a review, see Later, Baumeister, Campbell, Krueger and Vohs, 2003), anxiety (see Horwitz, 2010 for a summary) and extraversion (Dewaele, 2012), and general language aptitude measured using standardized tests (Swansea Language aptitude test used in Abrahamsson & Hyltenstam, 2008; Modern Language Aptitude Test used in Li, 2013). Current study adds to literature looking at how specific aspects of cognition relate to learning particular feature of language –lexical tones in Mandarin.

The most intriguing finding here was that the cognitive functions predicting learning differed for learners at different stages, with Musical Ability and Working Memory being more important for novice learners and Attention for later learners. This finding raises several new questions. Firstly, at what point in learning process do Musical Ability and Working Memory becomes less important and Attention becomes more important? While earlier work (Li & Francis, 2014) has highlighted the different attention allocation in language between native Mandarin speakers and native English speakers, preliminary work by Zou et al (2017) revealed L2 learners of Mandarin also tends to change their attention mechanisms in speech perception after learning Mandarin for at least 3 years. The current study shows that there are differences in learning processes between native English speakers with no experience in Mandarin and native English speakers who have learned Mandarin at the undergraduate level for just 18 months. How much exposure is necessary to lead to this change?

A second question is whether this shift of cognitive function would happen in a different pattern for learners at different ages. Recall that Sinkeviciute et al. (2019) found age differences in a L2 vocabulary learning study, where 7-year-old children showed greater difficulty

processing high variability materials compared with older children and adults. It might be possible that children might take longer in the “beginning” stage of learning so that the cognitive shift would take longer to appear (e.g. there may be no difference as for cognitive patterns between naïve children and those who has learned Mandarin for 18 months). It is also possible that different cognitive tasks might be predictive for children. For example, some of the simpler digit span tasks which were less relevant for adults might be more suitable for children.

Future work exploring these questions should potentially include more cognitive tasks than the current study and in particular a weakness of the current analysis is that only two Musical Ability tasks were used. This meant that effects of general music ability were unclear both in the composite measure used in Section 4.4 and in components extracted through PCA (Section 4.6) where it is grouped with other tasks that may share similar underlying mechanisms but not music ability per se (e.g. *Melody Memory* with working memory measures). In addition, as shown in Chan et al. (2006) and Machizawa and Driver (2011), in PCA the selection of factors entering the models is crucial for creating meaningful components. Although the current study has seen separate effects of Working Memory, Attention and Musical Ability between naïve participants and Mandarin learner participants, these patterns needs further replication with more/different tests to probe the underlying mechanisms (e.g. does Working Memory capacity contribute more to learning than processing speed).

Ultimately, research in this area may have practical implications: If we can identify the cognitive factors which are important in training, we may be able to adjust the training paradigm for learners with different cognitive profiles. This will be further discussed in Section 5.1.4.

### 5.1.3 Methodological Contribution

This thesis makes a methodological contribution in not only presenting null results but also introducing the use of Bayes Factors into phonetic training research which allow quantification of evidence from the H1 or the null.

As discussed above, very few published studies have reported HV versus LV comparisons, however it is possible that more have been conducted but these haven't been reported due to finding null results. There is increased concern in the field that researchers have been less willing to report null results (Franco, Malhotra & Simonovits, 2014) and journals have been less likely to accept articles reporting null results for publication (Ferguson & Heene, 2012). Evidence for this comes from the distribution of  $p$ -values in the published literature: psychological research normally uses  $p$  values as the main metric for inference - with a cut-off level of 0.05. Masicampo and Lalande (2012) reviewed 12 issues of three psychological journals before 2008. After examining the distribution of  $p$  values, they found – in all three journals - that  $p$  values were much more common between 0.45 and 0.50 than would be predicted by chance, suggesting a potential publication bias favoring statistical significance. Failing to report null results has the impact of preventing the full picture emerging in the area of language training research. The current thesis presents the null results for the benefit of HV input on generalisation and uses BF's to quantify evidence for the null. The studies presented in Chapters 2 & 3 has been published as Dong, Clayards, Brown & Wonnacott (2019) and along with Sinkeviciute et al. (2019) these are two of the first published studies in this area using Bayesian analysis.

The current thesis also contributes by applying Bayes Factor statistics when looking at cognitive factors predicting individual performance. As reported in Section 4.1, there are many previous studies looking for differences between a batteries of cognitive *ID measures* and then

reporting those which do and don't predict language-related behaviors. These papers generally use frequentist stats ( $p$ -values) but then draw conclusions about null results which are not valid – i.e. concluding that  $p > .05$  provides evidence *against* an effect. The current studies use statistics that can quantify the evidence for the null as well as H1. Although there are many places where the evidence was ambiguous for either H1 or the null, this approach at least makes this clear to the reader. Moreover, one advantage of using Bayes Factors is that – in principle where the evidence is ambiguous – I could keep increasing the sample pool and updating the measure of evidence until there is evidence for/ against the null. That is, unlike for  $p$ -values, there is no problem with optional stopping (Dienes, 2016). That said, for a lot of the ambiguous BFs found in the current results, additional analysis indicated that it would require a much larger sample – e.g. approximately seven times as big - to get the evidence for either the H1 or the null. Although this may seem infeasible, it may also indicate that it is necessary to increase sample sizes beyond what has been standard in psychological research. Nevertheless, there are many places in the data where there is a great deal of evidence for the effect in question (extremely large Bayes factors) even with the current sample. Using Bayesian statistics alongside traditional frequentist approach should be considered more in language training studies in the future.

Lastly, the second set of analyses run with PCA and regression has highlighted the difference between using single individual difference measures as predictors compared with extracted components loaded on multiple cognitive tasks. Overall, this new set of analyses did not provide a much clearer picture of the effects of cognitive factors. The results were mainly consistent with the results from models with separate ID measures as predictors however some of the components were hard to interpret. However, it did reveal some potential links between different cognitive tasks which may be worth exploring in future work. Importantly, the

analysis confirmed the general picture in which Attention is specifically important for current Mandarin learners whereas Working Memory is more important for naïve learners.

#### 5.1.4 *Future research direction*

As discussed in Chapter 3 (Section 3.4.3), one way to further examine the interaction between individual aptitude and training materials would be to conduct replications with high statistical power of the original studies by Perrachione et al. (2011) and Sadaka and McQueen (2014). Chapter 4 (Section 4.5.3) also suggested that, given the exploratory nature of the current work, a pre-registered replication of the current Study 3 is required. If the key findings of Study 3 were replicated successfully, an important future direction should bring together the two lines of research in this thesis, and add a low variability condition to the design in Study 3. Following previous work (Perrachione et al., 2011; Sadaka & McQueen, 2014), it is expected that participants with lower aptitude won't find LV materials as difficult as HV materials thus an interaction between certain *ID measures* and variability may be found. For naïve participants, an interaction between variability and Musical Ability would be predicted. If the Perrachione et al.'s hypothesis about aptitude is correct, not only will the correlation with musical ability be stronger for the HV materials than the LV materials, it may also be that participants with higher musical ability will benefit more from the HV training, while this is absent, or even reversed, for the LV training. For Mandarin learner participants, a similar pattern is expected for the interaction between variability and the measures of attention. Importantly, statistical values from the current study can be used to inform H1 for the BFs employed in this future study. Specifically, to estimate H1 for the interactions, the maximum difference expected would be if the HV condition replicates current effect sizes and the LV shows no correlation between performance and the relevant aptitude measures, if we continue to model H1 as a half normal distribution, the SD can be set as half of this maximum difference.

Again, it will be useful to pre-register this study, including the set of values which will be used as estimates of H1. If the results from this new study support these hypotheses, indicating that these cognitive individual difference measures do interact with variability, and if this differs for the learners at different stage of learning, then this could shed light on the important question of how different types of learning materials may be differently effective for different learners of different stages, which may have particularly practical implications. For example, for beginner Mandarin learners, for those with weaker musical ability, they may initially require materials to be less variable and explicitly focus on learning to hear the different pitch patterns. For later stage learners, for those who learn less due to having worse ability for attention shifting, we might also consider switching to materials with less variability such as in speakers and items, or could make other changes to make the paradigm less attention-demanding.

One interesting area for further exploration is whether there are differences in the extent to which training is effective for the six different Mandarin tone contrasts (and of course other tone contrasts in other languages). Recall from the introduction (Section 1.1) that the speech perception models PAM and SLM make interesting predictions about the learning of different tonal contrasts. The work of So and Best (2014, 2016) exploring the PAM theory has suggested that naïve participants may assimilate tonal information into existing English intonational categories and these categories may overlap. For example, both T1 and T4 may be assimilated into the category “Statement”, although with different category goodness, with T1 more like “Statement” than T4. This *category goodness assimilation* might make it easier to discriminate than identify these two tones: discriminating them can be done on the basis of how good an exemplar is compared with the assimilated L1 category, while identifying them relies on the ability to group both exemplars belonging to different L2 categories despite having assimilated to the same L1 category. However, for contrasts such as T1-T3, participants are likely to

assimilate them into different categories (T1: Statement; T3: question/uncertainty). Such two-category assimilation might happen, making it equally difficult for identification tasks and discrimination tasks. Thus, we might see that for the T1-T4 contrast, participants perform better in the discrimination task than the identification task, while for T1-T3 contrast, participants perform equally well in the two tasks. In terms of how these different contrasts might respond to training, SLM in particular emphasizes how repeated exposure and training can improve L2 perception by leading to category dissimilation and the creation of a new category (e.g. between T1, T3 & T4 as suggested by Hao, 2014). In Hao (2014), he found that there was no training effect for T2 only, which may imply that this tone is more involved in the intonational use of English. However this needs further exploration. It is also possible that the inherent differences between different tonal contrasts, as highlighted by these theories, might mean that they benefit differently from different levels of variability in training and/or are more/less susceptible to effects of individual differences (although these theories of L2 speech perception have not specifically considered either the role of multiple talker input or individual differences per se). In theory, the experiments reported in this thesis have relevant data to address this question. However, in all experiments reported in this thesis, all analyses have used as their DV a general measure of learning across all of the tones, and while factors for tone contrasts were not included in the models, I did *not* actually inspect the effect of different tones/tonal contrasts specifically, and did not look at how these contrasts interacted with *test-session* and *variability-condition*. This approach was taken in order to avoid having models with too many degrees of freedom and looking at lots of underpowered effects which increases the chance of Type 1 error. However, all of the data has been made available to other researchers online<sup>17</sup>. Exploratory analyses could be undertaken, looking for any effects specifically of tones/tonal

---

<sup>17</sup> ([https://osf.io/j6s7w/?view\\_only=497e0e8ee7ff4e7387984690eafd4b5a](https://osf.io/j6s7w/?view_only=497e0e8ee7ff4e7387984690eafd4b5a)).

contrasts. For example, do some contrasts show greater learning effects than others (suggested by the interaction between trials testing different tonal contrasts, and test-session)? Do any benefit more from variability in training (looking for interaction between trials testing different contrasts, test-session and variability-condition)? Any such exploratory analyses could then be followed up in pre-registered, confirmatory research which was properly powered to look at these effects. Such follow up work could potentially be important in terms of the practical applied goals of this thesis (e.g. showing if training programs should be made to focus more on some contrasts than others, or if different types of training were required for different contrasts). It might also shed further light on the mechanisms underpinning development of L2 lexical tones. Although this hasn't been the key focus of the current thesis, this is relevant for theoretical work in L2 speech perception and may have implications for the development of theories such as the PAM and SLM.

Another important direction is that future experiments should be designed with increased ecological validity. The current work moved in this direction by not only using real world stimuli but also including all four tones in training. There is evidence in the literature that training with the full set of phonetic features may be important. Two studies by Nishi and Kewley-Port (2007, 2008) found that it was more efficient to train Japanese speakers with all nine English monophthong vowels than just training three of them. Also, if they started the training with full set and narrowed it down to the subset, the training was equally effective as using full set. In particular, those who were trained with the full set could generalise to new items and the training effect was sustained for a longer period of time. Their explanation was that when learners created new phonetic categories, if only part of the vowel system was presented, then a set of categories may be formed such that it is hard to then adjust the boundaries to accommodate vowels that are encountered later. Thus, training English speakers with all four Mandarin tones, as opposed to two or three, may be important in establishing the

correct representation of tones, even if it makes training initially harder. The ecological validity may also be improved through other adjustments. For example, the current training paradigm trains monosyllabic words in isolation, while for natural Mandarin in daily use, the duration and the pitch of a tone also varies based on the surrounding syllables (e.g. the 5<sup>th</sup> tone, see section 1.2). Adding in this type of variability would certainly make training more difficult there is an important question as to whether it is better to train with monosyllables first, and then deal with this complexity, or train with more complicated materials to begin with. Future study should compare training with sentences versus isolated words, as well as different orderings of the two types of training. A further important question will be whether the benefits of these different training set-ups interact with cognitive individual differences.

## 6. References

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4), 433-459.
- Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in second language acquisition*, 30(4), 481-509.  
DOI: <https://doi.org/10.1017/S027226310808073X>
- Ahola, K., Vilkki, J., & Servo, A. (1996). Frontal tests do not detect frontal infarctions after ruptured intracranial aneurysm. *Brain and Cognition*, 31(1), 1-16. <https://doi.org/10.1006/brcg.1996.0021>
- Alais, D., Morrone, C., & Burr, D. (2006). Separate attentional resources for vision and audition. *Proceedings of the Royal Society B: Biological Sciences*, 273(1592), 1339-1345.  
<https://doi.org/10.1098/rspb.2005.3420>
- 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.
- Alshangiti, W., & Evans, B. G. (2014, May). Investigating the domain-specificity of phonetic training for second language 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*.
- Amitay, S., Irwin, A., & Moore, D. R. (2006). Discrimination learning induced by training with identical stimuli. *Nature neuroscience*, 9(11), 1446. <https://doi.org/10.1038/nn1787>
- Andrés, P., & Van der Linden, M. (2000). Age-related differences in supervisory attentional system functions. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, 55(6), P373-P380. <https://doi.org/10.1093/geronb/55.6.P373>
- Andrews, S. (2012). Individual differences in skilled visual word recognition and reading. *Visual word recognition*, 2, 151-172.
- Apfelbaum, K. S., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, 35(6), 1105-1138.  
<https://doi.org/10.1111/j.1551-6709.2011.01181.x>
- Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, 28, 403-450.  
<https://doi.org/10.1146/annurev.neuro.28.061604.135709>
- Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In *Psychology of learning and motivation* (Vol. 2, pp. 89-195). Academic Press.  
[https://doi.org/10.1016/S0079-7421\(08\)60422-3](https://doi.org/10.1016/S0079-7421(08)60422-3)
- Audacity Team. (2015). Audacity (Version 2.1.1). *Computer Program*. Retrieved May, 2015, from <http://audacityteam.org/>
- Awh, E., Vogel, E. K., & Oh, S. H. (2006). Interactions between attention and working memory. *Neuroscience*, 139(1), 201-208. <https://doi.org/10.1016/j.neuroscience.2005.08.023>

- 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>
- Baddeley, A. (1996). Exploring the central executive. *The Quarterly Journal of Experimental Psychology Section A*, 49(1), 5-28. <https://doi.org/10.1080/713755608>
- Baddeley, A. (2000). The episodic buffer: a new component of working memory?. *Trends in cognitive sciences*, 4(11), 417-423. [https://doi.org/10.1016/S1364-6613\(00\)01538-2](https://doi.org/10.1016/S1364-6613(00)01538-2)
- Baddeley, A. (2003). Working memory and language: An overview. *Journal of communication disorders*, 36(3), 189-208. [https://doi.org/10.1016/S0021-9924\(03\)00019-4](https://doi.org/10.1016/S0021-9924(03)00019-4)
- Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual review of psychology*, 63, 1-29. <https://doi.org/10.1146/annurev-psych-120710-100422>
- Baddeley, A. D. (1966a). Short-term Memory for Word Sequences as a Function of Acoustic, Semantic and Formal Similarity. *Quarterly Journal of Experimental Psychology*, 18(4), 362–365.
- Baddeley, A. D. (1966b). The influence of acoustic and semantic similarity on long-term memory for word sequences. *The Quarterly journal of experimental psychology*, 18(4), 302-309. <https://doi.org/10.1080/14640746608400047>
- Baddeley, A. D. (1986). *Working memory*. Oxford: Oxford University Press.
- Baddeley, A. D., & Hitch, G. (1974). Working memory. In *Psychology of learning and motivation* (Vol. 8, pp. 47-89). Academic press.
- Baddeley, A. D., & Warrington, E. K. (1970). Amnesia and the distinction between long-and short-term memory. *Journal of verbal learning and verbal behavior*, 9(2), 176-189.
- Baddeley, A. D., Allen, R. J., & Hitch, G. J. (2011). Binding in visual working memory: The role of the episodic buffer. *Neuropsychologia*, 49(6), 1393-1400. <https://doi.org/10.1016/j.neuropsychologia.2010.12.042>
- Baddeley, A. D., Hitch, G. J., & Allen, R. J. (2009). Working memory and binding in sentence recall. *Journal of Memory and Language*, 61(3), 438-456. <https://doi.org/10.1016/j.jml.2009.05.004>
- Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6), 575-589. [https://doi.org/10.1016/S0022-5371\(75\)80045-4](https://doi.org/10.1016/S0022-5371(75)80045-4)
- Baddeley, A., & Wilson, B. A. (2002). Prose recall and amnesia: Implications for the structure of working memory. *Neuropsychologia*, 40(10), 1737-1743. [https://doi.org/10.1016/S0028-3932\(01\)00146-4](https://doi.org/10.1016/S0028-3932(01)00146-4)
- Baddeley, A., Chincotta, D., Stafford, L., & Turk, D. (2002). Is the word length effect in STM entirely attributable to output delay? Evidence from serial recognition. *The Quarterly Journal of Experimental Psychology Section A*, 55(2), 353-369. <https://doi.org/10.1080/02724980143000523>
- Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological review*, 105(1), 158. DOI:[10.1037/0033-295x.105.1.158](https://doi.org/10.1037/0033-295x.105.1.158)

- Baddeley, A.D., & Lieberman, K. (1980). Spatial working memory. In R. S. Nickerson (Ed.), *Attention and performance VIII* (pp. 521–539). Hillsdale, NJ: Lawrence Erlbaum.
- Bak, T. H., Vega-Mendoza, M., & Sorace, A. (2014). Never too late? An advantage on tests of auditory attention extends to late bilinguals. *Frontiers in psychology, 5*, 485.  
<https://doi.org/10.3389/fpsyg.2014.00485>
- Bakermans-Kranenburg, M. J., & Van Ijzendoorn, M. H. (2011). Differential susceptibility to rearing environment depending on dopamine-related genes: New evidence and a meta-analysis. *Development and psychopathology, 23*(1), 39-52.  
DOI: <https://doi.org/10.1017/S0954579410000635>
- 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>
- Barr, R. A., & Giambra, L. M. (1990). Age-related decrement in auditory selective attention. *Psychology and Aging, 5*(4), 597. DOI:[10.1037//0882-7974.5.4.597](https://doi.org/10.1037//0882-7974.5.4.597)
- Barton, B., & Neville-Barton, P. (2004). Undergraduate mathematics learning in English by speakers of other languages. In *Topic Study Group 25 at the 10th International Congress on Mathematics Education, July*.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0–5. 2013.
- Baumeister, R. F., Campbell, J. D., Krueger, J. I., & Vohs, K. D. (2003). Does high self-esteem cause better performance, interpersonal success, happiness, or healthier lifestyles?. *Psychological science in the public interest, 4*(1), 1-44. <https://doi.org/10.1111/1529-1006.01431>
- Beane, M., & Marrocco, R. T. (2004). Norepinephrine and acetylcholine mediation of the components of reflexive attention: implications for attention deficit disorders. *Progress in neurobiology, 74*(3), 167-181. <https://doi.org/10.1016/j.pneurobio.2004.09.001>
- Benson, N., Hulac, D. M., & Kranzler, J. H. (2010). Independent examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): what does the WAIS-IV measure?. *Psychological Assessment, 22*(1), 121.
- Bergsleithner, J. M. (2011). The role of noticing and working memory capacity in L2 oral performance. *Organon, 26*(51). DOI: <https://doi.org/10.22456/2238-8915.28841>.
- Best, C. T. (1994). The Emergence of Native-Language Phonological Influences in Infants: A Perceptual Assimilation Model. In J. C. Goodman & H. C. Nusbaum (Eds.), *The Development of Speech Perception: The Transition from Speech Sounds to Spoken Words 1* (Vol. 167, pp. 167–224). Cambridge, MA: MIT Press.

- Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). York Press.
- Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. *Language experience in second language speech learning: In honor of James Emil Flege, 1334*, 1-47.
- Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *Journal for the Association for Research in Otolaryngology*, 8(2), 294-304. doi: [10.1007/s10162-007-0073-z](https://doi.org/10.1007/s10162-007-0073-z)
- Bhatara, A., Yeung, H. H., & Nazzi, T. (2015). Foreign language learning in French speakers is associated with rhythm perception, but not with melody perception. *Journal of Experimental Psychology: Human Perception and Performance*, 41(2), 277. <http://dx.doi.org/10.1037/a0038736>
- Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child development*, 70(3), 636-644. <https://doi.org/10.1111/1467-8624.00046>
- Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental science*, 7(3), 325-339. <https://doi.org/10.1111/j.1467-7687.2004.00351.x>
- Bialystok, E., Craik, F. I., Klein, R., & Viswanathan, M. (2004). Bilingualism, aging, and cognitive control: evidence from the Simon task. *Psychology and aging*, 19(2), 290. <http://dx.doi.org/10.1037/0882-7974.19.2.290>
- Bidet-Caulet, A., Voisin, J., Bertrand, O., & Fonlupt, P. (2005). Listening to a walking human activates the temporal biological motion area. *Neuroimage*, 28(1), 132-139. <https://doi.org/10.1016/j.neuroimage.2005.06.018>
- Blascovich, J., & Tomaka, J. (1991). Measures of self-esteem. *Measures of personality and social psychological attitudes*, 1, 115-160.
- 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/>
- Bornstein, M. H., Tamis-LeMonda, C. S., & Haynes, O. M. (1999). First words in the second year: Continuity, stability, and models of concurrent and predictive correspondence in vocabulary and verbal responsiveness across age and context. *Infant Behavior and Development*, 22(1), 65-85. [https://doi.org/10.1016/S0163-6383\(99\)80006-X](https://doi.org/10.1016/S0163-6383(99)80006-X)
- Bouvrie, J. V. (2009). *Hierarchical learning: Theory with applications in speech and vision* (Doctoral dissertation, Massachusetts Institute of Technology).
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729. <https://doi.org/10.1016/j.cognition.2007.04.005>
- Bradlow, A. R., & Pisoni, D. B. (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., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l: Long-term retention of learning in perception and production. *Perception & psychophysics*, 61(5), 977-985. <https://doi.org/10.3758/BF03206911>
- Brown, H. D. (2000). *Principles of language learning and teaching*. White Plains, NY: Longman.
- Brown, J. (1958). Some tests of the decay theory of immediate memory. *Quarterly Journal of Experimental Psychology*, 10(1), 12-21. <https://doi.org/10.1080/17470215808416249>
- Bygate, M., Swain, M., & Skehan, P. (2013). *Researching pedagogic tasks: Second language learning, teaching, and testing*. London UK: Routledge.
- Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2012). Does first language maintenance hamper nativelikeness in a second language?: A study of ultimate attainment in early bilinguals. *Studies in Second Language Acquisition*, 34(2), 215-241. DOI: <https://doi.org/10.1017/S0272263112000034>
- Cao, J. (2016). *语言的韵律与语音的变化*. Beijing: China social sciences press.
- Carlet, A., & Cebrian, J. (2015). Identification vs. discrimination training: Learning effects for trained and untrained sounds. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow.
- Carretti, B., Borella, E., Cornoldi, C., & De Beni, R. (2009). Role of working memory in explaining the performance of individuals with specific reading comprehension difficulties: A meta-analysis. *Learning and individual differences*, 19(2), 246-251. <https://doi.org/10.1016/j.lindif.2008.10.002>
- Carroll, J. B., & Sapon, S. M. (1959). *Modern language aptitude test*. San Antonio, TX, US: Psychological Corporation.
- Carter, C. S., & Krug, M. K. (2012). Dynamic cognitive control and frontal-cingulate interactions. *Cognitive neuroscience of attention*, 2, 89-98.
- Chan, E., Skehan, P., & Gong, G. (2011). Working memory, phonemic coding ability and foreign language aptitude: Potential for construction of specific language aptitude tests—the case of Cantonese. *Ilha do Desterro: A Journal of English Language, Literatures in English and Cultural Studies*, (60), 45-73. DOI: <https://doi.org/10.5007/2175-8026.2011n60p045>
- Chan, R. C., Hoosain, R., & Lee, T. M. (2002). Reliability and validity of the Cantonese version of the test of everyday attention among normal Hong Kong Chinese: a preliminary report. *Clinical Rehabilitation*, 16(8), 900-909. <https://doi.org/10.1191/0269215502cr574oa>
- Chan, R. C., Lai, M. K., & Robertson, I. H. (2006). Latent structure of the Test of Everyday Attention in a non-clinical Chinese sample. *Archives of clinical neuropsychology*, 21(5), 477-485.
- Chandrasekaran, B., Kraus, N., & Wong, P. C. (2011). Human inferior colliculus activity relates to individual differences in spoken language learning. *Journal of neurophysiology*, 107(5), 1325-1336. <https://doi.org/10.1152/jn.00923.2011>
- Chandrasekaran, B., Sampath, P. D., & Wong, P. C. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America*, 128(1), 456-465. <https://doi.org/10.1121/1.3445785>

- Chen, H., Qian, Y. H., Zhou, X. Q., & Guo, M. H. (2010). Study of three-dimensional speech chart by time-frequency analysis in Chinese Mandarin Monosyllabic word phonemes. *南方医科大学报 Journal of Southern Medical University*, 30(8), 1805-1809.
- Chen, Z. Y., Cowell, P. E., Varley, R., & Wang, Y. C. (2009). A cross-language study of verbal and visuospatial working memory span. *Journal of Clinical and Experimental Neuropsychology*, 31(4), 385-391. <https://doi.org/10.1080/13803390802195195>
- Cheng, C. C. (2011). *A synchronic phonology of Mandarin Chinese* (Vol. 4). Walter de Gruyter.
- Cheung, H. (1996). Nonword span as a unique predictor of second-language vocabulary language. *Developmental Psychology*, 32(5), 867. <https://doi.org/10.1037/0012-1649.32.5.867>
- Chinese Academy of Social Sciences (2012). *中国语言地图集(第2版):汉语方言卷 [Language Atlas of China (2nd edition): Chinese dialect volume]*. Beijing: The Commercial Press, ISBN 978-7-100-07054-6.
- ChinesePod. 2019. ChinesePod.com. Retrieved from: <http://chinesepod.com/tools/pronunciation/section/17>
- 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](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.
- Collette, F., & Van der Linden, M. (2002). Brain imaging of the central executive component of working memory. *Neuroscience & Biobehavioral Reviews*, 26(2), 105-125. [https://doi.org/10.1016/S0149-7634\(01\)00063-X](https://doi.org/10.1016/S0149-7634(01)00063-X)
- Conboy, B. T., Sommerville, J. A., & Kuhl, P. K. (2008). Cognitive control factors in speech perception at 11 months. *Developmental psychology*, 44(5), 1505. doi: [10.1037/a0012975](https://doi.org/10.1037/a0012975)
- Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British journal of psychology*, 55(4), 429-432. <https://doi.org/10.1111/j.2044-8295.1964.tb00928.x>
- Corsi, P. M. (1972). Human memory and the medial temporal region of the brain. *Dissertation Abstracts International*, 34, 819B.
- Cowan, N. (2016). *Working Memory Capacity: Classic Edition*. Routledge.
- Creel, S. C., Weng, M., Fu, G., Heyman, G. D., & Lee, K. (2018). Speaking a tone language enhances musical pitch perception in 3–5-year-olds. *Developmental science*, 21(1). <https://doi.org/10.1111/desc.12503>
- Crowe, S. F. (2000). Does the letter number sequencing task measure anything more than digit span?. *Assessment*, 7(2), 113-117. <https://doi.org/10.1177/107319110000700202>
- Davis, M. (1994). Folk music psychology. *The psychologist*, 7(12), 537.

- Dehaene, S., & Changeux, J. P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, 70(2), 200-227. <https://doi.org/10.1016/j.neuron.2011.03.018>
- Della Sala, S., Gray, C., Baddeley, A., Allamano, N., & Wilson, L. (1999). Pattern span: a tool for unwelding visuo-spatial memory. *Neuropsychologia*, 37(10), 1189-1199. [https://doi.org/10.1016/S0028-3932\(98\)00159-6](https://doi.org/10.1016/S0028-3932(98)00159-6)
- Delogu, F., Lampis, G., & Belardinelli, M. O. (2010). From melody to lexical tone: Musical ability enhances specific aspects of foreign language perception. *European Journal of Cognitive Psychology*, 22(1), 46-61. <https://doi.org/10.1080/09541440802708136>
- Delogu, F., Lampis, G. and Olivetti Belardinelli, M. 2006. Music-to-language transfer effect: May melodic ability improve learning of tonal languages by native nontonal speakers?. *Cognitive Processing*, 7: 203–207. [10.1007/s10339-006-0146-7](https://doi.org/10.1007/s10339-006-0146-7)
- Dewaele, J. M. (2012). Personality in second language acquisition. *The Encyclopedia of Applied Linguistics*.
- Dewaele, J.-M. (2007). Predicting language learners' grades in the L1, L2, L3 and L4: The effect of some psychological and sociocognitive variables. *International Journal of Multilingualism*, 4(3), 169–97. <https://doi.org/10.2167/ijm080.0>
- Dewaele, J.-M. (2002) Individual differences in L2 fluency: the effect of neurobiological correlates. In: Cook, V. (ed.) *Portraits of the L2 user*. Bristol, UK: Multilingual Matters, pp. 219-250. ISBN 9781853595837.
- Díaz, B., Baus, C., Escera, C., Costa, A., & Sebastián-Gallés, N. (2008). Brain potentials to native phoneme discrimination reveal the origin of individual differences in learning the sounds of a second language. *Proceedings of the National Academy of Sciences*, 105(42), 16083-16088. <https://doi.org/10.1073/pnas.0805022105>
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.
- Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in psychology*, 5, 781. <https://doi.org/10.3389/fpsyg.2014.00781>
- Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In M. Overgaard (Ed.), *Behavioural methods in consciousness research* (pp. 199–220). Oxford: Oxford University Press.
- Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, 72, 78-89. <https://doi.org/10.1016/j.jmp.2015.10.003>
- Dienes, Z., Coulton, S., & Heather, N. (2018). Using Bayes factors to evaluate evidence for no effect: examples from the SIPS project. *Addiction*, 113(2), 240-246. <https://doi.org/10.1111/add.14002>
- Dong, H., Clayards, M., Brown, H., & Wonnacott, E. (2019). The effects of high versus low talker variability and individual aptitude on phonetic training of Mandarin lexical tones. *PeerJ*, 7, e7191. DOI: [10.7717/peerj.7191](https://doi.org/10.7717/peerj.7191)
- Dong, Y., Tsubota, Y., & Dantsuji, M. (2013, November). Difficulties in perception and pronunciation of Mandarin Chinese disyllabic word tone acquisition: A study of some Japanese university students.

In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)* (pp. 143-152).

Dörnyei, Z. (1994). Motivation and motivating in the foreign language classroom. *The modern language journal*, 78(3), 273-284. DOI: 10.2307/330107 DOI: 10.2307/330107

Dosenbach, N. U., Fair, D. A., Cohen, A. L., Schlaggar, B. L., & Petersen, S. E. (2008). A dual-networks architecture of top-down control. *Trends in cognitive sciences*, 12(3), 99-105.  
<https://doi.org/10.1016/j.tics.2008.01.001>

Duanmu, S. (2007). *The phonology of standard Chinese*. OUP Oxford.

Egi, T., Fujii, A., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. *Individual differences and instructed language learning*, 181-209.

Ellis, N. C. (2006). Cognitive perspectives on SLA: The associative-cognitive CREED. *Aila Review*, 19(1), 100-121. DOI: <https://doi.org/10.1075/aila.19.08ell>

Ellis, R. (1994). Factors in the incidental acquisition of second language vocabulary from oral input: A review essay. *Applied Language Learning*, 5(1), 1-32.

Ellis, R. (2004). Individual Differences in Second Language Learning. *The handbook of applied linguistics*, 525.

Eme, Cecilia & Odinye, Sunny. (2008). Phonology of standard Chinese and Igbo: Implications for Igbo students learning Chinese. *NKOA Journal*, 1.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological review*, 100(3), 363. [10.1037/0033-295X.100.3.363](https://doi.org/10.1037/0033-295X.100.3.363)

Erlam, R. (2005). Language aptitude and its relationship to instructional effectiveness in second language acquisition. *Language Teaching Research*, 9(2), 147-171.  
<https://doi.org/10.1191/1362168805lr161oa>

Facoetti, A., Trussardi, A. N., Ruffino, M., Lorusso, M. L., Cattaneo, C., Galli, R., ... & Zorzi, M. (2010). Multisensory spatial attention deficits are predictive of phonological decoding skills in developmental dyslexia. *Journal of cognitive neuroscience*, 22(5), 1011-1025.  
<https://doi.org/10.1162/jocn.2009.21232>

Fan, J., McCandliss, B. D., Fossella, J., Flombaum, J. I., & Posner, M. I. (2005). The activation of attentional networks. *Neuroimage*, 26(2), 471-479.  
<https://doi.org/10.1016/j.neuroimage.2005.02.004>

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of cognitive neuroscience*, 14(3), 340-347.  
<https://doi.org/10.1162/089892902317361886>

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, 7(6), 555-561.  
<https://doi.org/10.1177/1745691612459059>

Field, A. (2009). *Discovering statistics using SPSS (2nd edition ed.)*. London: Sage.

- Fitch, W. T. (2005). The evolution of language: a comparative review. *Biology and philosophy*, 20(2-3), 193-203.
- Flege, J. 1987. The production of 'new' and 'similar' phones in a foreign language: *Evidence for the effect of equivalence classification*. *Journal of Phonetics*, 15, 47-65.
- Flege, J. E. (1995). Second Language Speech Learning: Theory, Findings, and Problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233–277). Timonium, MD: York Press.
- Flege, J. E. (2007). Language contact in bilingualism: Phonetic system interactions. In J. Cole & J. Hualde (eds) *Laboratory phonology 9*, 353-382. Berlin: Mouton de Gruyter.
- Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English/l/and/l/accurately. *Language and Speech*, 38, 25-55.  
<https://doi.org/10.1177/002383099503800102>
- Flege, J. E. 1995. Two procedures for training a novel second language phonetic contrast. *Applied Psycholinguistics*, 16, 425–442 DOI: <https://doi.org/10.1017/S0142716400066029>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502-1505. DOI: 10.1126/science.1255484
- Galligan, L. (2001). Possible effects of English-Chinese language differences on the processing of mathematical text: A review. *Mathematics Education Research Journal*, 13(2), 112-132.  
<https://doi.org/10.1007/BF03217102>
- Gardner, R. C. (2014). Attitudes and motivation in second language learning. In *Bilingualism, multiculturalism, and second language learning* (pp. 63-84). Psychology Press.
- Gathercole, S. E. (1995). Is nonword repetition a test of phonological memory or long-term knowledge? It all depends on the nonwords. *Memory and Cognition*, 23, 83–94.  
<https://doi.org/10.3758/BF03210559>
- Gathercole, S. E., Tiffany, C., Briscoe, J., Thorn, A., & ALSPAC team. (2005). Developmental consequences of poor phonological short-term memory function in childhood: A longitudinal study. *Journal of child Psychology and Psychiatry*, 46(6), 598-611. <https://doi.org/10.1111/j.1469-7610.2004.00379.x>
- Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental psychology*, 28(5), 887.
- Gathercole, S.E., Baddeley, A.D. (1989). Development of vocabulary in children and short-term phonological memory. *Journal of Memory and Language*, 28, 200-213
- Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: bridging selective attention and working memory. *Trends in cognitive sciences*, 16(2), 129-135. <https://doi.org/10.1016/j.tics.2011.11.014>
- George, E. M., & Coch, D. (2011). Music training and working memory: an ERP study. *Neuropsychologia*, 49(5), 1083-1094. <https://doi.org/10.1016/j.neuropsychologia.2011.02.001>
- Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory?. *Trends in cognitive sciences*, 10(6), 278-285. <https://doi.org/10.1016/j.tics.2006.04.008>

- 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](https://doi.org/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>
- Gibson, C., Folley, B. S., & Park, S. (2009). Enhanced divergent thinking and creativity in musicians: A behavioral and near-infrared spectroscopy study. *Brain and cognition*, 69(1), 162-169. <https://doi.org/10.1016/j.bandc.2008.07.009>
- Gilbert, J. P. (1981). Motoric music skill development in young children: A longitudinal investigation. *Psychology of Music*, 9(1), 21-25.
- Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(1), 152.
- Gordon, E.E. (1979). *Primary measures of music audiation*. Chicago, IL: GIA.
- Gordon, E.E. (1982). *Intermediate measures of music audiation*. Chicago, IL: GIA.
- Gordon, E.E. (1989). *Advanced measures of music audiation*. Chicago, IL: GIA.
- Gottfried, T. L., & Ouyang, G. Y. H. (2005). Production of Mandarin tone contrasts by musicians and non-musicians. *The Journal of the Acoustical Society of America*, 118(3), 2025-2025. <https://doi.org/10.1121/1.4785767>
- Gottfried, T. L., & Riester, D. (2000). Relation of pitch glide perception and Mandarin tone identification. *Journal of the Acoustical Society of America*, 108(5), 2604. <https://doi.org/10.1121/1.4743698>
- Gottfried, T. L., Staby, A. M., & Ziemer, C. J. (2004). Musical experience and Mandarin tone discrimination and imitation. *The Journal of the Acoustical Society of America*, 115(5), 2545-2545. <https://doi.org/10.1121/1.4783674>
- Grant, L. D., & Weissman, D. H. (2017). An attentional mechanism for minimizing cross-modal distraction. *Acta psychologica*, 174, 9-16. <https://doi.org/10.1016/j.actpsy.2017.01.003>
- Graybiel, A. M. (1977). Direct and indirect precolomotor pathways of the brainstem: an autoradiographic study of the pontine reticular formation in the cat. *Journal of Comparative Neurology*, 175(1), 37-78. <https://doi.org/10.1002/cne.901750105>
- Grenon, I., Kubota, M., & Sheppard, C. (2019). The creation of a new vowel category by adult learners after adaptive phonetic training. *Journal of Phonetics*, 72, 17-34. <https://doi.org/10.1016/j.wocn.2018.10.005>
- Gruber, O., & Goschke, T. (2004). Executive control emerging from dynamic interactions between brain systems mediating language, working memory and attentional processes. *Acta psychologica*, 115(2-3), 105-121. <https://doi.org/10.1016/j.actpsy.2003.12.003>
- Hallam, S. (1998). The predictors of achievement and drop out in instrumental tuition. *Psychology of Music*, 26(2), 116-132. <https://doi.org/10.1177/0305735698262002>

- Hallam, S. (2004). How important is practicing as a predictor of learning outcomes in instrumental music?. *learning*, 4, 1-76.
- Hallam, S. (2010). 21st century conceptions of musical ability. *Psychology of Music*, 38(3), 308-330. <https://doi.org/10.1177/0305735609351922>
- Hallam, S., & Prince, V. (2003). Conceptions of musical ability. *Research Studies in Music Education*, 20(1), 2-22. <https://doi.org/10.1177/1321103X030200010101>
- Hallé, P. A., Chang, Y. C., & Best, C. T. (2004). Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs. French listeners. *Journal of phonetics*, 32(3), 395-421. [https://doi.org/10.1016/S0095-4470\(03\)00016-0](https://doi.org/10.1016/S0095-4470(03)00016-0)
- Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1), 23.
- Han, Y., & Ginsburg, H. P. (2001). Chinese and English mathematics language: The relation between linguistic clarity and mathematics performance. *Mathematical Thinking and Learning*, 3(2-3), 201-220. <https://doi.org/10.1080/10986065.2001.9679973>
- Hanel, P. H., & Vione, K. C. (2016). Do student samples provide an accurate estimate of the general public?. *PloS one*, 11(12). doi: [10.1371/journal.pone.0168354](https://doi.org/10.1371/journal.pone.0168354)
- Hao, Y. C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. *Journal of phonetics*, 40(2), 269-279. <https://doi.org/10.1016/j.wocn.2011.11.001>
- Hao, Y. C. (2014). The application of the Speech Learning Model to the L2 acquisition of Mandarin tones. In *Fourth International Symposium on Tonal Aspects of Languages*.
- Hari, R., & Renvall, H. (2001). Impaired processing of rapid stimulus sequences in dyslexia. *Trends in cognitive sciences*, 5(12), 525-532. [https://doi.org/10.1016/S1364-6613\(00\)01801-5](https://doi.org/10.1016/S1364-6613(00)01801-5)
- Haroutounian, J. (2000). Perspectives of musical talent: A study of identification criteria and procedures. *High Ability Studies*, 11(2), 137-160. <https://doi.org/10.1080/13598130020001197>
- Hay, J., Drager, K., & Thomas, B. (2013). Using nonsense words to investigate vowel merger. *English Language & Linguistics*, 17(2), 241-269.
- Hazan, V., & Kim, Y. H. (2010). Can we predict who will benefit from computer-based phonetic training?. In *Second Language Studies: Acquisition, Learning, Education and Technology*.
- Heinrich, A., Schneider, B. A., & Craik, F. I. (2008). Investigating the influence of continuous babble on auditory short-term memory performance. *The Quarterly Journal of Experimental Psychology*, 61(5), 735-751. <https://doi.org/10.1080/17470210701402372>
- Henry, P. J. (2008). Student sampling as a theoretical problem. *Psychological Inquiry*, 19(2), 114-126. <https://doi.org/10.1080/10478400802049951>
- Hill, B. D., Elliott, E. M., Shelton, J. T., Pella, R. D., O'Jile, J. R., & Gouvier, W. D. (2010). Can we improve the clinical assessment of working memory? An evaluation of the Wechsler Adult Intelligence Scale—Third Edition using a working memory criterion construct. *Journal of Clinical and Experimental Neuropsychology*, 32(3), 315-323.

- Hoch, L., & Tillmann, B. (2012). Shared structural and temporal integration resources for music and arithmetic processing. *Acta Psychologica*, 140(3), 230-235.  
<https://doi.org/10.1016/j.actpsy.2012.03.008>
- Horwitz, E. K. (2010). Foreign and second language anxiety. *Language Teaching*, 43(2), 154-167.  
 DOI: <https://doi.org/10.1017/S026144480999036X>
- Hötting, K., Rösler, F., & Röder, B. (2003). Crossmodal and intermodal attention modulate event-related brain potentials to tactile and auditory stimuli. *Experimental Brain Research*, 148(1), 26-37.  
<https://doi.org/10.1111/desc.12503>
- Hu, J. (2015). *现代汉语基础(第二版)*. Peking University Press.
- Hurley, C. G. (1995). Student motivations for beginning and continuing/discontinuing string music instruction. *The Quarterly Journal of Music Teaching and Learning*, 6(1), 44-55.
- Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12), 1489-1506. [https://doi.org/10.1016/S0042-6989\(99\)00163-7](https://doi.org/10.1016/S0042-6989(99)00163-7)
- Iversen, J.R., Patel, A.D. (2008). The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. In: *Proceedings of the 10th International Conference on Music Perception and Cognition (ICMPC 10)*. Sapporo: Japan, 465-468.
- Iverson, P., & Evans, B. G. (2007). Learning English vowels with different first-language vowel systems: Perception of formant targets, formant movement, and duration. *The Journal of the Acoustical Society of America*, 122(5), 2842-2854. <https://doi.org/10.1121/1.2783198>
- Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, 126(2), 866-877. <https://doi.org/10.1121/1.3148196>
- 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>
- Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.
- Jentschke, S., & Koelsch, S. (2009). Musical training modulates the development of syntax processing in children. *Neuroimage*, 47(2), 735-744. <https://doi.org/10.1016/j.neuroimage.2009.04.090>
- 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>

- Johnson, J. A., & Zatorre, R. J. (2005). Attention to simultaneous unrelated auditory and visual events: behavioral and neural correlates. *Cerebral Cortex*, *15*(10), 1609-1620.  
<https://doi.org/10.1093/cercor/bhi039>
- Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, *29*(1), 1-23. <https://doi.org/10.1006/cogp.1995.1010>
- Kaye, A. S. (2002). Vowels and Consonants: An Introduction to the Sounds of Languages. *Language*, *78*(2), 361-362. DOI: <https://doi.org/10.1353/lan.2007.0015>
- Kayser, C., Petkov, C. I., Augath, M., & Logothetis, N. K. (2007). Functional imaging reveals visual modulation of specific fields in auditory cortex. *Journal of Neuroscience*, *27*(8), 1824-1835. DOI: <https://doi.org/10.1523/JNEUROSCI.4737-06.2007>
- Kayser, C., Petkov, C. I., Lippert, M., & Logothetis, N. K. (2005). Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology*, *15*(21), 1943-1947.  
<https://doi.org/10.1016/j.cub.2005.09.040>
- Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and Mandarin. *The Journal of the Acoustical Society of America*, *132*(2), 1050-1060.  
<https://doi.org/10.1121/1.4730893>
- Keizer, K., & Kuypers, H. G. J. M. (1989). Distribution of corticospinal neurons with collaterals to the lower brain stem reticular formation in monkey (*Macaca fascicularis*). *Experimental Brain Research*, *74*(2), 311-318. <https://doi.org/10.1007/BF00248864>
- Kessels, R. P., Van Zandvoort, M. J., Postma, A., Kappelle, L. J., & De Haan, E. H. (2000). The Corsi block-tapping task: standardization and normative data. *Applied neuropsychology*, *7*(4), 252-258.  
[https://doi.org/10.1207/S15324826AN0704\\_8](https://doi.org/10.1207/S15324826AN0704_8)
- Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help?. *Language, cognition and neuroscience*, *34*(1), 43-68.  
<https://doi.org/10.1080/23273798.2018.1500698>
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, *122*(2), 148.  
doi: [10.1037/a0038695](https://doi.org/10.1037/a0038695)
- Koelsch, S., Gunter, T. C., Cramon, D. Y. V., Zysset, S., Lohmann, G., & Friederici, A. D. (2002). Bach speaks: a cortical "language-network" serves the processing of music. *Neuroimage*, *17*(2), 956-966.  
doi:10.1006/nimg.2002.1154.
- Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and cognition*, *11*(2), 261-271. DOI: <https://doi.org/10.1017/S1366728908003416>
- Krashen, S. (1981). Second language acquisition. *Second Language Learning*, *3*(7), 19-39.
- Kraus, N., Strait, D. L., & Parbery-Clark, A. (2012). Cognitive factors shape brain networks for auditory skills: spotlight on auditory working memory. *Annals of the New York Academy of Sciences*, *1252*(1), 100-107. doi: [10.1111/j.1749-6632.2012.06463.x](https://doi.org/10.1111/j.1749-6632.2012.06463.x)

- Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research*, 25(1), 161-168. <https://doi.org/10.1016/j.cogbrainres.2005.05.004>
- Kroll, N. E., Parks, T., Parkinson, S. R., Bieber, S. L., & Johnson, A. L. (1970). Short-term memory while shadowing: recall of visually and of aurally presented letters. *Journal of Experimental Psychology*, 85(2), 220.
- Krumhansl, C. L., & Keil, F. C. (1982). Acquisition of the hierarchy of tonal functions in music. *Memory & cognition*, 10(3), 243-251. <https://doi.org/10.3758/BF03197636>
- Lallier, M., Tainturier, M. J., Dering, B., Donnadieu, S., Valdois, S., & Thierry, G. (2010). Behavioral and ERP evidence for a modal sluggish attentional shifting in developmental dyslexia. *Neuropsychologia*, 48(14), 4125-4135. <https://doi.org/10.1016/j.neuropsychologia.2010.09.027>
- Lalonde, C. E., & Werker, J. F. (1995). Cognitive influences on cross-language speech perception in infancy. *Infant Behavior and Development*, 18(4), 459-475. [https://doi.org/10.1016/0163-6383\(95\)90035-7](https://doi.org/10.1016/0163-6383(95)90035-7)
- 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*, 33(1), 159-174. DOI: 10.2307/2529310
- Larsen, H., van der Zwaluw, C. S., Overbeek, G., Granic, I., Franke, B., & Engels, R. C. (2010). A variable-number-of-tandem-repeats polymorphism in the dopamine D4 receptor gene affects social adaptation of alcohol use: Investigation of a gene-environment interaction. *Psychological Science*, 21(8), 1064-1068. <https://doi.org/10.1177/0956797610376654>
- Leary, M. R., Tate, E. B., Adams, C. E., Batts Allen, A., & Hancock, J. (2007). Self-compassion and reactions to unpleasant self-relevant events: The implications of treating oneself kindly. *Journal of personality and social psychology*, 92(5), 887. <http://dx.doi.org/10.1037/0022-3514.92.5.887>
- Lee, C. Y., & Hung, T. H. (2008). Identification of Mandarin tones by English-speaking musicians and nonmusicians. *The Journal of the Acoustical Society of America*, 124(5), 3235-3248. <https://doi.org/10.1121/1.2990713>
- Lee, M., & Faber, R. J. (2007). Effects of product placement in on-line games on brand memory: A perspective of the limited-capacity model of attention. *Journal of advertising*, 36(4), 75-90. <https://doi.org/10.2753/JOA0091-3367360406>
- Lee, W. S., & Zee, E. (2003). Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1), 109-112. DOI: <https://doi.org/10.1017/S0025100303001208>
- 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>

- Leon, A. C., & Heo, M. (2009). Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational statistics & data analysis*, 53(3), 603-608.
- Lev-Ari, S., & Peperkamp, S. (2014). The influence of inhibitory skill on phonological representations in production and perception. *Journal of Phonetics*, 47, 36-46.  
<https://doi.org/10.1016/j.wocn.2014.09.001>
- Levy, E. S. (2009). Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French. *The Journal of the Acoustical Society of America*, 125(2), 1138-1152. <https://doi.org/10.1121/1.3050256>
- Li, M., & DeKeyser, R. (2017). Perception practice, production practice, and musical ability in L2 Mandarin tone-word learning. *Studies in Second Language Acquisition*, 39(4), 593-620.  
DOI: <https://doi.org/10.1017/S0272263116000358>
- Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *The Modern Language Journal*, 97(3), 634-654. <https://doi.org/10.1111/j.1540-4781.2013.12030.x>
- Lieberman, A. M. (1996). *Speech: A special code*. MIT press.
- Lieberman, M. D., & Rosenthal, R. (2001). Why introverts can't always tell who likes them: Multitasking and nonverbal decoding. *Journal of Personality and Social Psychology*, 80, 294-310.  
[10.1037/0022-3514.80.2.294](https://doi.org/10.1037/0022-3514.80.2.294)
- Lin, H. & Wang, Q. (2007). Mandarin rhythm: An acoustic study. *Journal of chinese language and computing*, 17(3), 127-140.
- Lin, M., & Francis, A. L. (2014). Effects of language experience and expectations on attention to consonants and tones in English and Mandarin Chinese. *The Journal of the Acoustical Society of America*, 136(5), 2827-2838. <https://doi.org/10.1121/1.4898047>
- Lindner, A., Iyer, A., Kagan, I., & Andersen, R. A. (2010). Human posterior parietal cortex plans where to reach and what to avoid. *Journal of Neuroscience*, 30(35), 11715-11725.  
<https://doi.org/10.1523/JNEUROSCI.2849-09.2010>
- Liu, M. (2012). Predicting effects of personality traits, self-esteem, language class risk-taking and sociability on Chinese university EFL learners' performance in English. *Journal of Second Language Teaching and Research*, 1(1), 30-57. ISSN 2045-4031
- Liu, Z. (2013). *汉语拼音经典方案选评*. Beijing: Beijing Language and Culture University Press.
- 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>
- Locke, J. L. (1989). Babbling and early speech: Continuity and individual differences. *First Language*, 9(6), 191-205. <https://doi.org/10.1177/014272378900900606>
- 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>

- Logie, R. H. (1986). Visuo-spatial processing in working memory. *The Quarterly Journal of Experimental Psychology Section A*, 38(2), 229-247. <https://doi.org/10.1080/14640748608401596>
- Logie, R. H. (2011). The Functional Organization and Capacity Limits of Working Memory. *Current Directions in Psychological Science*, 20(4), 240–245. <https://doi.org/10.1177/09637214111415340>
- Logie, R. H. (2016). Retiring the central executive. *The Quarterly Journal of Experimental Psychology*, 69(10), 2093-2109. <https://doi.org/10.1080/17470218.2015.1136657>
- Luinge, M. R., Post, W. J., Wit, H. P., & Goorhuis-Brouwer, S. M. (2006). The ordering of milestones in language development for children from 1 to 6 years of age. *Journal of Speech, Language, and Hearing Research*, 49(5), 923-940. [https://doi.org/10.1044/1092-4388\(2006/067\)](https://doi.org/10.1044/1092-4388(2006/067))
- Lukmani, Y. M. (1972). Motivation to learn and language proficiency. *Language learning*, 22(2), 261-273. <https://doi.org/10.1111/j.1467-1770.1972.tb00087.x>
- Lundin, R. W. (1953). *An objective psychology of music*. Oxford, England: Ronald Press.
- Lynn, R., Wilson, R. G., & Gault, A. (1989). Simple musical tests as measures of Spearman's g. *Personality and Individual Differences*, 10(1), 25-28. [https://doi.org/10.1016/0191-8869\(89\)90173-6](https://doi.org/10.1016/0191-8869(89)90173-6)
- Machizawa, M. G., & Driver, J. (2011). Principal component analysis of behavioural individual differences suggests that particular aspects of visual working memory may relate to specific aspects of attention. *Neuropsychologia*, 49(6), 1518-1526.
- MacIntyre, P. D., & Gardner, R. C. (1991). Methods and results in the study of anxiety and language learning: A review of the literature. *Language learning*, 41(1), 85-117. <https://doi.org/10.1111/j.1467-1770.1991.tb00677.x>
- MacIntyre, P. D., Clément, R., & Noels, K. A. (2007). Affective variables, attitude and personality in context. *Handbook of French applied linguistics*, 270-298.
- MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing willingness to communicate in a L2: A situational model of L2 confidence and affiliation. *The Modern Language Journal*, 82(4), 545-562. <https://doi.org/10.1111/j.1540-4781.1998.tb05543.x>
- Magne, C., Schon, D. and Besson, M. (2006). Musician children detect pitch violations in both music and language better than non-musician children: Behavioral and electrophysiological approaches. *Journal of Cognitive Neurosciences*, 18: 199–211. <https://doi.org/10.1162/jocn.2006.18.2.199>
- 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>
- Marie, C., Delogu, F., Lampis, G., Belardinelli, M. O., & Besson, M. (2011). Influence of musical expertise on segmental and tonal processing in Mandarin Chinese. *Journal of Cognitive Neuroscience*, 23(10), 2701-2715. <https://doi.org/10.1162/jocn.2010.21585>
- Marques, C., Moreno, S., Luís Castro, S., & Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience*, 19(9), 1453-1463. <https://doi.org/10.1162/jocn.2007.19.9.1453>

- Martinussen, R., & Tannock, R. (2006). Working memory impairments in children with attention-deficit hyperactivity disorder with and without comorbid language learning disorders. *Journal of clinical and experimental neuropsychology*, 28(7), 1073-1094. <https://doi.org/10.1080/13803390500205700>
- Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, 65(11), 2271-2279. <https://doi.org/10.1080/17470218.2012.711335>
- Mattock, K., & Burnham, D. (2006). Chinese and English infants' tone perception: Evidence for perceptual reorganization. *Infancy*, 10(3), 241-265. [https://doi.org/10.1207/s15327078in1003\\_3](https://doi.org/10.1207/s15327078in1003_3)
- 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>
- Mattys, S. L., Baddeley, A., & Trenkic, D. (2018). Is the superior verbal memory span of Mandarin speakers due to faster rehearsal?. *Attention, Perception, & Psychophysics*, 79(3), 945-963. <https://doi.org/10.3758/s13414-017-1283->
- Maxwell, O., Baker, B., Bundgaard-Nielsen, R., & Fletcher, J. (2015). A comparison of the acoustics of nonsense and real word stimuli: Coronal stops in Bengali. *Proceedings of the meeting of the International Congress of Phonetic Sciences*, Glasgow, UK. Retrieved from <http://www.icphs2015.info/>
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, 70(6), 487. <http://dx.doi.org/10.1037/a0039400>
- McCabe, D. P., Roediger III, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. *Neuropsychology*, 24(2), 222. doi: [10.1037/a0017619](https://doi.org/10.1037/a0017619)
- 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>
- McPherson, G. E. (1995). Five aspects of musical performance and their correlates. *Bulletin of the council for research in music education*, 115-121. <https://www.jstor.org/stable/40318774>
- McPherson, G. E., & Renwick, J. M. (2011). Self-regulation and mastery of musical skills. *Handbook of self-regulation of learning and performance*, 234-248.
- Meara, P. (2005). LLAMA language aptitude tests: The manual. *Swansea: Lognostics*.
- Meda, S. A., Stevens, M. C., Potenza, M. N., Pittman, B., Gueorguieva, R., Andrews, M. M., ... & Pearlson, G. D. (2009). Investigating the behavioral and self-report constructs of impulsivity domains using principal component analysis. *Behavioural pharmacology*, 20(5-6), 390. doi: [10.1097/FBP.0b013e32833113a3](https://doi.org/10.1097/FBP.0b013e32833113a3)
- Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing research*, 219(1-2), 36-47. <https://doi.org/10.1016/j.heares.2006.05.004>
- Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2), 81.

- Milovanov, R., Huotilainen, M., Välimäki, V., Esquef, P. A., & Tervaniemi, M. (2008). Musical aptitude and second language pronunciation skills in school-aged children: Neural and behavioral evidence. *Brain research*, 1194, 81-89.  
<https://doi.org/10.1016/j.brainres.2007.11.042>
- Misyak, J. B., Christiansen, M. H., & Bruce Tomblin, J. (2010). Sequential expectations: The role of prediction-based learning in language. *Topics in Cognitive Science*, 2(1), 138-153.  
<https://doi.org/10.1111/j.1756-8765.2009.01072.x>
- Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.
- Moore, D. G., Burland, K., & Davidson, J. W. (2003). The social context of musical success: A developmental account. *British Journal of Psychology*, 94(4), 529-549.  
<https://doi.org/10.1348/000712603322503088>
- Moreno, S., Bialystok, E., Barac, R., Schellenberg, E. G., Cepeda, N. J., & Chau, T. (2011). Short-term music training enhances verbal intelligence and executive function. *Psychological science*, 22(11), 1425-1433. <https://doi.org/10.1177/0956797611416999>
- Moreno, S., Marques, C., Santos, A., Santos, M., Castro, S. L., & Besson, M. (2008). Musical training influences linguistic abilities in 8-year-old children: more evidence for brain plasticity. *Cerebral Cortex*, 19(3), 712-723. <https://doi.org/10.1093/cercor/bhn120>
- Morgan, J. L., & Demuth, K. (Eds.). (2014). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. Psychology Press.
- Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental science*, 10(6), 719-726. <https://doi.org/10.1111/j.1467-7687.2007.00623.x>
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: an index for assessing musical sophistication in the general population. *PloS one*, 9(2), e89642. <https://doi.org/10.1371/journal.pone.0101091>
- Müllensiefen, D., Harrison, P., Caprini, F., & FancThe t, A. (2015). Investigating the importance of self-theories of intelligence and musicality for students' academic and musical achievement. *Frontiers in Psychology*, 6 (1702). <https://doi.org/10.3389/fpsyg.2015.01702>
- Musacchia, G., Strait, D., & Kraus, N. (2008). Relationships between behavior, brainstem and cortical encoding of seen and heard speech in musicians and non-musicians. *Hearing research*, 241(1-2), 34-42. <https://doi.org/10.1016/j.heares.2008.04.013>
- National Bureau of Statistics of China, 2010. China Statistics Year Book, 2010. Retrieved from <http://www.stats.gov.cn/tjsj/ndsj/2010/indexch.htm>.
- Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive science*, 14(1), 11-28.  
[https://doi.org/10.1016/0364-0213\(90\)90024-Q](https://doi.org/10.1016/0364-0213(90)90024-Q)
- Ni, Y. Wang, X. (1992). 英语国家学生学习汉语语音难点分析. *Chinese Language Learning*, 2, 47-60.
- Nikolin, S., Martin, D. M., Loo, C., & Lauf, S. (2018). Effects of high-definition transcranial direct current stimulation (HD-tDCS) of the intraparietal sulcus and dorsolateral prefrontal cortex on

- working memory and divided attention. *Frontiers in integrative neuroscience*, 12, 64.  
<https://doi.org/10.3389/fnint.2018.00064>
- Noels, K., Clément, R., & Pelletier, L. (2001). Intrinsic, extrinsic, and integrative orientations of French Canadian learners of English. *Canadian Modern Language Review*, 57(3), 424-442.  
<https://doi.org/10.3138/cmlr.57.3.424>
- Norrix, L. W., Plante, E., & Vance, R. (2006). Auditory–visual speech integration by adults with and without language-learning disabilities. *Journal of Communication Disorders*, 39(1), 22-36.  
<https://doi.org/10.1016/j.jcomdis.2005.05.003>
- Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, 115(1), 39.
- Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. *Speech perception, production and linguistic structure*, (pp. 113-134). Amsterdam, Netherlands: IOS press.
- O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second language oral production by adult learners. *Applied Psycholinguistics*, 27(3), 377-402. DOI: <https://doi.org/10.1017/S0142716406060322>
- O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior research methods, instruments, & computers*, 32(3), 396-402.
- Ong, G., Sewell, D. K., Weekes, B., McKague, M., & Abutalebi, J. (2017). A diffusion model approach to analysing the bilingual advantage for the Flanker task: The role of attentional control processes. *Journal of Neurolinguistics*, 43, 28-38. <https://doi.org/10.1016/j.jneuroling.2016.08.002>
- Ou, J., & Law, S. P. (2017). Cognitive basis of individual differences in speech perception, production and representations: The role of domain general attentional switching. *Psychonomic bulletin & review*, 22(6), 1725-1732. <https://doi.org/10.3758/s13414-017-1283-z>
- Ou, J., Law, S. P., & Fung, R. (2015). Relationship between individual differences in speech processing and cognitive functions. *Psychonomic bulletin & review*, 22(6), 1725-1732.  
<https://doi.org/10.3758/s13423-015-0839-y>
- Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive psychology*, 66(2), 232-258.  
<https://doi.org/10.1016/j.cogpsych.2012.12.002>
- Palmer, C., & Pfordresher, P. Q. (2003). Incremental planning in sequence production. *Psychological Review*, 110(4), 683. DOI: [10.1037/0033-295X.110.4.683](https://doi.org/10.1037/0033-295X.110.4.683)
- Parbery-Clark, A., Tierney, A., Strait, D. L., & Kraus, N. (2012). Musicians have fine-tuned neural distinction of speech syllables. *Neuroscience*, 219, 111-119.  
<https://doi.org/10.1016/j.neuroscience.2012.05.042>
- Patel, A. D. (2003). Language, music, syntax and the brain. *Nature neuroscience*, 6(7), 674.  
**DOI:** [10.1038/nn1082](https://doi.org/10.1038/nn1082)
- Patel, A. D. (2006). Musical rhythm, linguistic rhythm, and human evolution. *Music Perception: An Interdisciplinary Journal*, 24(1), 99-104. DOI: 10.1525/mp.2006.24.1.99

- Patel, A. D. (2010). *Music, language, and the brain*. Oxford university press.
- Pawlak, M. (2012). The dynamic nature of motivation in language learning: A classroom perspective. *Studies in Second Language Learning and Teaching*, 2(2), 249-278. DOI: [10.14746/ssl.2012.2.2.7](https://doi.org/10.14746/ssl.2012.2.2.7)
- Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., & Sams, M. (2006). Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. *Human Brain Mapping*, 27(6), 471-477. <https://doi.org/10.1002/hbm.20190>
- Peng, G. (2006). Temporal and tonal aspects of Chinese syllables: A corpus-based comparative study of Mandarin and Cantonese. *Journal of Chinese Linguistics*, 34(1), 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 design. *The Journal of the Acoustical Society of America*, 130, 461-472. <https://doi.org/10.1371/journal.pone.0089642>
- Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: the CDP+ model of reading aloud. *Psychological review*, 114(2), 273. <https://doi.org/10.1037/0033-295X.114.2.273>
- Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of experimental psychology*, 58(3), 193.
- Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. *Perception & Psychophysics*, 16(2), 283-290. <https://doi.org/10.3758/BF03203943>
- Phillips, W. A., & Baddeley, A. D. (1971). Reaction time and short-term visual memory. *Psychonomic Science*, 22(2), 73-74. <https://doi.org/10.3758/BF03332500>
- Pike, K. L. (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press.
- Pisoni, D. B., & Cleary, M. (2003). Measures of working memory span and verbal rehearsal speed in deaf children after cochlear implantation. *Ear and hearing*, 24(1 Suppl), 106S. doi: [10.1097/01.AUD.0000051692.05140.8E](https://doi.org/10.1097/01.AUD.0000051692.05140.8E)
- Posner, M. I. (1978). *Chronometric explorations of mind*. Lawrence Erlbaum.
- Posner, M. I., & Keele, S. W. (1967). Decay of visual information from a single letter. *Science*, 158(3797), 137-139.
- Posner, M. I., & Konick, A. F. (1966). Short-term retention of visual and kinesthetic information. *Organizational Behavior and Human Performance*, 1(1), 71-86.
- Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual review of neuroscience*, 13(1), 25-42.
- Posner, M. I., Rothbart, M. K., Sheese, B. E., & Voelker, P. (2012). Control networks and neuromodulators of early development. *Developmental psychology*, 48(3), 827. doi: [10.1037/a0025530](https://doi.org/10.1037/a0025530)
- 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](http://www.r-project.org). Accessed September 2017.
- Ramscar, M., & Baayen, H. (2013). Production, comprehension, and synthesis: a communicative perspective on language. *Frontiers in psychology*, 4, 233. <https://doi.org/10.3389/fpsyg.2013.00233>
- Ramscar, M., Yarlett, D., Dye, M., Denny, K., & Thorpe, K. (2010). The effects of feature-label-order and their implications for symbolic learning. *Cognitive science*, 34(6), 909-957. <https://doi.org/10.1111/j.1551-6709.2009.01092.x>
- Reber, A. S., Walkenfeld, F. F., & Hernstadt, R. (1991). Implicit and explicit learning: Individual differences and IQ. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 17(5), 888-896. DOI: [10.1037//0278-7393.17.5.888](https://doi.org/10.1037//0278-7393.17.5.888)
- Reder, F., Marec-Breton, N., Gombert, J. E., & Demont, E. (2013). Second-language learners' advantage in metalinguistic awareness: A question of languages' characteristics. *British Journal of Educational Psychology*, 83(4), 686-702. <https://doi.org/10.1111/bjep.12003>
- Revelle, W. (2019). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 1.9.12, <https://CRAN.R-project.org/package=psych>.
- Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). The test of everyday attention (TEA). *Bury St. Edmunds, UK: Thames Valley Test Company*, 197-221.
- Robinson, P. (1997). Individual differences and the fundamental similarity of implicit and explicit adult second language learning. *Language learning*, 47(1), 45-99. <https://doi.org/10.1111/0023-8333.21997002>
- Robinson, P., Mackey, A., Gass, S. M., & Schmidt, R. (2012). Attention and awareness in second language acquisition. *The Routledge handbook of second language acquisition*, 247-267.
- Rodriguez, C. X. (1998). Children's perception, production, and description of musical expression. *Journal of Research in Music Education*, 46(1), 48-61. <https://doi.org/10.2307/3345759>
- Rubin, D. C. (1995). *Memory in oral traditions: The cognitive psychology of epic, ballads, and counting-out rhymes*. Oxford University Press on Demand.
- Rueda, M. R., Rothbart, M. K., McCandliss, B. D., Saccomanno, L., & Posner, M. I. (2005). Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences*, 102(41), 14931-14936. <https://doi.org/10.1073/pnas.0506897102>
- Ryan, J. J., & Paolo, A. M. (2001). Exploratory factor analysis of the WAIS-III in a mixed patient sample. *Archives of Clinical Neuropsychology*, 16(2), 151-156.
- Sadakata, M., & McQueen, J. M. (2013). High stimulus variability in nonnative speech learning supports formation of abstract categories: Evidence from Japanese geminates. *The Journal of the Acoustical Society of America*, 134(2), 1324-1335. <https://doi.org/10.1121/1.4812767>

- Sadakata, M., & McQueen, J. M. (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>
- Scarborough, R. (2012). Lexical similarity and speech production: Neighborhoods for nonwords. *Lingua*, *122*(2), 164-176. <https://doi.org/10.1016/j.lingua.2011.06.006>
- Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J. L., & Besson, M. (2010). Similar cerebral networks in language, music and song perception. *Neuroimage*, *51*(1), 450-461. <https://doi.org/10.1016/j.neuroimage.2010.02.023>
- Schulz, P., & Grimm, A. (2018). The age factor revisited: Timing in acquisition interacts with age of onset in bilingual acquisition. *Frontiers in psychology*, *9*. doi: [10.3389/fpsyg.2018.02732](https://doi.org/10.3389/fpsyg.2018.02732)
- Seashore, C. E., Lewis, D., & Saetveit, J. G. (1956). *Seashore measures of musical talents*. Oxford, England: Psychological Corp.
- Service, E. & Kohonen, V. (1995). Is the relation between phonological memory and foreign language learning accounted for by vocabulary acquisition? *Applied Psycholinguistics*, *16*(2), 155-172. DOI: <https://doi.org/10.1017/S0142716400007062>
- Shallice, T., & Warrington, E. K. (1970). Independent functioning of verbal memory stores: A neuropsychological study. *The Quarterly journal of experimental psychology*, *22*(2), 261-273. <https://doi.org/10.1080/0033557043000203>
- Shi, R. (2014). Functional morphemes and early language acquisition. *Child Development Perspectives*, *8*(1), 6-11. <https://doi.org/10.1111/cdep.12052>
- Shomstein, S., & Yantis, S. (2004). Control of attention shifts between vision and audition in human cortex. *Journal of neuroscience*, *24*(47), 10702-10706. DOI: <https://doi.org/10.1523/JNEUROSCI.2939-04.2004>
- Shuter-Dyson, R. (1999). Musical ability. In *The psychology of music* (pp. 627-651). Academic Press.
- Sinkeviciute, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, 1-26. DOI: <https://doi.org/10.1017/S0272263119000263>
- Sjölander, A., & Vansteelandt, S. (2019). Frequentist versus Bayesian approaches to multiple testing. *European journal of epidemiology*, 1-13. doi: [10.1007/s10654-019-00517-2](https://doi.org/10.1007/s10654-019-00517-2)
- Skehan, P. (1991). Individual differences in second language learning. *Studies in second language acquisition*, *13*(2), 275-298. DOI: <https://doi.org/10.1017/S0272263100009979>
- Skehan, P. (2002). Theorizing and updating aptitude. In P. Robinson (Ed.), *Individual differences in instructed language learning* (pp. 69-93). Amsterdam: John Benjamins.
- Slevc, L. R., & Miyake, A. (2006). Individual differences in second-language proficiency: Does musical ability matter?. *Psychological Science*, *17*(8), 675-681. <https://doi.org/10.1111/j.1467-9280.2006.01765.x>

- Slevc, L. R., Rosenberg, J. C., & Patel, A. D. (2009). Making psycholinguistics musical: self-paced reading time evidence for shared processing of linguistic and musical syntax. *Psychonomic bulletin & review*, 16(2), 374-381. <https://doi.org/10.3758/16.2.374>
- Sloboda, J. A., Davidson, J. W., Howe, M. J., & Moore, D. G. (1996). The role of practice in the development of performing musicians. *British journal of psychology*, 87(2), 287-309. <https://doi.org/10.1111/j.2044-8295.1996.tb02591.x>
- Sloboda, J. A., Davidson, J. W., Howe, M. J., & Murphy, P. (2000). Is everyone musical. *Learners, learning and assessment*, 46-57.
- Smyth, M. M., & Pendleton, L. R. (1990). Space and movement in working memory. *The Quarterly Journal of Experimental Psychology Section A*, 42(2), 291-304. <https://doi.org/10.1080/14640749008401223>
- So, C. K., & Best, C. T. (2008). Do English speakers assimilate Mandarin tones to English prosodic categories?. In *Ninth Annual Conference of the International Speech Communication Association. INTERSPEECH-2008*, 1120.
- So, C. K., & Best, C. T. (2011). Categorizing Mandarin tones into listeners' native prosodic categories: The role of phonetic properties. *Poznań Studies in Contemporary Linguistics PSiCL*, 47, 133. DOI: <https://doi.org/10.2478/psicl-2011-0011>
- So, C. K., & Best, C. T. (2014). Phonetic influences on English and French listeners' assimilation of mandarin tones to native prosodic categories. *Studies in Second Language Acquisition*, 36(2), 195-221. DOI: <https://doi.org/10.1017/S0272263114000047>
- 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>
- Speciale, G., Ellis, N. C. & Bywater, T. (2004). Phonological sequence learning and short-term store capacity determine second language vocabulary acquisition. *Applied Psycholinguistics*, 25, 293-321. DOI: <https://doi.org/10.1017/S0142716404001146>
- Spielmann, G., & Radnofsky, M. L. (2001). Learning language under tension: New directions from a qualitative study. *The Modern Language Journal*, 85(2), 259-278. <https://doi.org/10.1111/0026-7902.00108>
- Spolsky, B. (2000). Anniversary article. Language motivation revisited. *Applied linguistics*, 21(2), 157-169. <https://doi.org/10.1093/applin/21.2.157>
- Statista. (2019). The most spoken languages worldwide. Retrieved on 15, October, 2019, from <https://www.statista.com/statistics/266808/the-most-spoken-languages-worldwide/>.
- Stigler, J. W., Lee, S. Y., & Stevenson, H. W. (1986). Digit memory in Chinese and English: Evidence for a temporally limited store. *Cognition*, 23(1), 1-20. [https://doi.org/10.1016/0010-0277\(86\)90051-X](https://doi.org/10.1016/0010-0277(86)90051-X)

- Strait, D. L., Kraus, N., Parbery-Clark, A., & Ashley, R. (2010). Musical experience shapes top-down auditory mechanisms: evidence from masking and auditory attention performance. *Hearing research*, 261(1-2), 22-29. <https://doi.org/10.1016/j.heares.2009.12.021>
- 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>
- Strange, W., & Shafer, V. L. (2008). Speech perception in second language learners: The re-education of selective perception. *Phonology and second language acquisition* (pp.153-192). Amsterdam, Netherland: John Benjamins Publishing Company.
- Strange, W., Weber, A., Levy, E. S., Shafiro, V., Hisagi, M., & Nishi, K. (2007). Acoustic variability within and across German, French, and American English vowels: Phonetic context effects. *The Journal of the Acoustical Society of America*, 122(2), 1111-1129. <https://doi.org/10.1121/1.2749716>
- Sturm, W., & Willmes, K. (2001). On the functional neuroanatomy of intrinsic and phasic alertness. *Neuroimage*, 14(1), S76-S84. <https://doi.org/10.1006/nimg.2001.0839>
- Stuss, D. T., & Knight, R. T. (Eds.). (2013). *Principles of frontal lobe function*. Oxford University Press.
- Sun, C. (2006). *Chinese: A Linguistic Introduction*. Cambridge: Cambridge University Press.
- Sussman, E., Winkler, I., & Schröger, E. (2003). Top-down control over involuntary attention switching in the auditory modality. *Psychonomic bulletin & review*, 10(3), 630-637. <https://doi.org/10.3758/BF03196525>
- Swaminathan, S., & Gopinath, J. K. (2013). Music training and second-language English comprehension and vocabulary skills in Indian children. *Psychological Studies*, 58(2), 164-170. <https://doi.org/10.1007/s12646-013-0180-3>
- Tagarelli, K. M., Mota, M. B., & Rebuschat, P. (2011). The role of working memory in implicit and explicit language learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 33, No. 33).
- Talamini, F., Altoè, G., Carretti, B., & Grassi, M. (2017). Musicians have better memory than nonmusicians: A meta-analysis. *PloS one*, 12(10), e0186773. <https://doi.org/10.1371/journal.pone.0186773>
- Tallal, P., & Piercy, M. (1973). Defects of non-verbal auditory perception in children with developmental aphasia. *Nature*, 241(5390), 468. <https://doi.org/10.1038/241468a0>
- Tallal, P., Miller, S. L., Bedi, G., Byma, G., Wang, X., Nagarajan, S. S., ... & Merzenich, M. M. (1996). Language comprehension in language-learning impaired children improved with acoustically modified speech. *Science*, 271(5245), 81-84. DOI: 10.1126/science.271.5245.81
- Tan, L. H., Liu, H. L., Perfetti, C. A., Spinks, J. A., Fox, P. T., & Gao, J. H. (2001). The neural system underlying Chinese logograph reading. *Neuroimage*, 13(5), 836-846. <https://doi.org/10.1006/nimg.2001.0749>
- Tang, Y. Y., Ma, Y., Fan, Y., Feng, H., Wang, J., Feng, S., ... & Zhang, Y. (2009). Central and autonomic nervous system interaction is altered by short-term meditation. *Proceedings of the national Academy of Sciences*, 106(22), 8865-8870. <https://doi.org/10.1073/pnas.0904031106>

- Tao, D., Deng, R., Jiang, Y., Galvin III, J. J., Fu, Q. J., & Chen, B. (2014). Contribution of auditory working memory to speech understanding in mandarin-speaking cochlear implant users. *PLoS one*, 9(6), e99096. <https://doi.org/10.1371/journal.pone.0099096>
- Thompson, K. G., Biscoe, K. L., & Sato, T. R. (2005). Neuronal basis of covert spatial attention in the frontal eye field. *Journal of Neuroscience*, 25(41), 9479-9487. DOI: <https://doi.org/10.1523/JNEUROSCI.0741-05.2005>
- Trahiotis, C., Bernstein, L. R., Buell, T. N., & Spektor, Z. (1990). On the use of adaptive procedures in binaural experiments. *The Journal of the Acoustical Society of America*, 87(3), 1359-1361. <https://doi.org/10.1121/1.399513>
- Trainor, L. J. (2005). Are there critical periods for musical development?. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 46(3), 262-278. <https://doi.org/10.1002/dev.20059>
- Trehub, S. E. (2001). Musical predispositions in infancy. *Annals of the New York academy of sciences*, 930(1), 1-16. DOI: [10.1111/j.1749-6632.2001.tb05721.x](https://doi.org/10.1111/j.1749-6632.2001.tb05721.x)
- Trehub, S. E., Bull, D., & Thorpe, L. A. (1984). Infants' perception of melodies: The role of melodic contour. *Child development*, 821-830. DOI: 10.2307/1130133
- Turnbull, O. H., Denis, M., Mellet, E., Ghaëm, O., & Carey, D. P. (2012). The processing of visuo-spatial information: Neuropsychological and neuroimaging investigations. In *Imagery, Language and Visuo-Spatial Thinking* (pp. 97-124). Psychology Press.
- Tye, C., Asherson, P., Ashwood, K. L., Azadi, B., Bolton, P., & McLoughlin, G. (2014). Attention and inhibition in children with ASD, ADHD and co-morbid ASD+ ADHD: an event-related potential study. *Psychological medicine*, 44(5), 1101-1116. DOI: <https://doi.org/10.1017/S0033291713001049>
- UKCISA. (2019). International student statistics: UK higher education. Retrieved on October 15, 2019, from <https://www.ukcisa.org.uk/Research--Policy/Statistics/International-student-statistics-UK-higher-education>.
- Vallar, G., & Baddeley, A. D. (1984). Fractionation of working memory: Neuropsychological evidence for a phonological short-term store. *Journal of Verbal Learning and Verbal Behavior*, 23(2), 151-161. [https://doi.org/10.1016/S0022-5371\(84\)90104-X](https://doi.org/10.1016/S0022-5371(84)90104-X)
- Van den Noort, M., Struys, E., Bosch, P., Jaswetz, L., Perriard, B., Yeo, S., ... & Lim, S. (2019). Does the bilingual advantage in cognitive control exist and if so, what are its modulating factors? A systematic review. *Behavioral Sciences*, 9(3), 27. <https://doi.org/10.3390/bs9030027>
- Van Rinsveld, A., Brunner, M., Landerl, K., Schiltz, C., & Ugen, S. (2015). The relation between language and arithmetic in bilinguals: insights from different stages of language acquisition. *Frontiers in psychology*, 6, 265. <https://doi.org/10.3389/fpsyg.2015.00265>
- Van Rinsveld, A., Schiltz, C., Brunner, M., Landerl, K., & Ugen, S. (2016). Solving arithmetic problems in first and second language: Does the language context matter?. *Learning and instruction*, 42, 72-82. <https://doi.org/10.1016/j.learninstruc.2016.01.003>
- VanPatten, B., & Smith, M. (2015). Aptitude as grammatical sensitivity and the initial stages of learning Japanese as a L2: Parametric variation and case marking. *Studies in Second Language Acquisition*, 37(1), 135-165. DOI: <https://doi.org/10.1017/S0272263114000345>

- Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151-175. <https://doi.org/10.1016/j.jml.2018.07.004>
- van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12. <https://doi.org/10.1016/j.jesp.2016.03.004>
- Veldre, A., & Andrews, S. (2014). Lexical quality and eye movements: Individual differences in the perceptual span of skilled adult readers. *The Quarterly Journal of Experimental Psychology*, *67*(4), 703-727. <https://doi.org/10.1080/17470218.2013.826258>
- Vidyasagar, T. R., & Pammer, K. (2010). Dyslexia: a deficit in visuo-spatial attention, not in phonological processing. *Trends in cognitive sciences*, *14*(2), 57-63. <https://doi.org/10.1016/j.tics.2009.12.003>
- Vispoel, W. P. (1993). The development and evaluation of a computerized adaptive test of tonal memory. *Journal of Research in Music Education*, *41*(2), 111-136. <https://doi.org/10.2307/3345403>
- Vogel, E. K., Woodman, G. F., & Luck, S. J. (2001). Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 92. <https://doi.org/10.1037/0096-1523.27.1.92>
- Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, *20*(3), 188-196. <https://doi.org/10.1016/j.lindif.2010.02.004>
- Wang, S. Q. (2019). China most popular destination for study abroad in Asia. Retrieved on 15 October, from <http://www.chinadaily.com.cn/a/201906/29/WS5d16fcc1a3103dbf1432afa5.html>.
- 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*(2), 1033-1043. <https://doi.org/10.1121/1.1531176>
- 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>
- Ward, P., Hodges, N. J., Williams, A. M., & Starkes, J. L. (2004). 11 Deliberate practice and expert performance. *Skill acquisition in sport: Research, theory and practice*, 231.
- Webster, P. R. (1988). New perspectives on music aptitude and achievement. *Psychomusicology: A Journal of Research in Music Cognition*, *7*(2), 177.
- Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised (WAIS—R): Manual*. San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (1997). *Wais-iii/wms-iii technical manual*. San Antonio, TX: The Psychological Corporation.

- Wechsler, D. (2008). Wechsler adult intelligence scale—Fourth Edition (WAIS—IV). *San Antonio, TX: NCS Pearson.*
- Weiss, L. G., Saklofske, D. H., Coalson, D., & Raiford, S. E. (Eds.). (2010). *WAIS-IV clinical use and interpretation: Scientist-practitioner perspectives.* Academic Press.
- Weissheimer, J., & Mota, M. B. (2011). Working memory capacity and the development of L2 speech production: A study of individual differences. In *Selected Proceedings of the 2010 Second Language Research Forum* (pp. 169-181).
- Weissman, D. H., Warner, L. M., & Woldorff, M. G. (2004). The neural mechanisms for minimizing cross-modal distraction. *Journal of Neuroscience, 24*(48), 10941-10949. DOI: <https://doi.org/10.1523/JNEUROSCI.3669-04.2004>
- Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development, 7*(1), 49-63. [https://doi.org/10.1016/S0163-6383\(84\)80022-3](https://doi.org/10.1016/S0163-6383(84)80022-3)
- Whalen, D. H., & Xu, Y. (1992). Information for Mandarin tones in the amplitude contour and in brief segments. *Phonetica, 49*(1), 25-47. <https://doi.org/10.1159/000261901>
- Wild-Wall, N., Falkenstein, M., & Gajewski, P. D. (2011). Age-related differences in working memory performance in a 2-back task. *Frontiers in psychology, 2*, 186. doi: [10.3389/fpsyg.2011.00186](https://doi.org/10.3389/fpsyg.2011.00186)
- Williamon, A., & Valentine, E. (2000). Quantity and quality of musical practice as predictors of performance quality. *British Journal of Psychology, 91*(3), 353-376. <https://doi.org/10.1348/000712600161871>
- Williams, M., & Burden, R. L. (1997). *Psychology for Language Teachers: A Social Constructivist Approach.* Cambridge University Press, 40 West 20th Street, New York, NY 10011-4211.
- Wilt, J., & Revelle, W. (2017). Extraversion. In T. A. Widiger (Ed.), *Oxford library of psychology. The Oxford handbook of the Five Factor Model* (pp. 57-81). New York, NY, US: Oxford University Press.
- Wing, H. D. (1968). *Tests of musical ability and appreciation: An investigation into the measurement, distribution, and development of musical capacity (2nd ed.).* London: Cambridge University Press.
- Wing, H. D. (1970). *Standardised tests of musical intelligence.* NFER-Nelson Publishing Company.
- Womelsdorf, T., Schoffelen, J. M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., & Fries, P. (2007). Modulation of neuronal interactions through neuronal synchronization. *Science, 316*(5831), 1609-1612. DOI: [10.1126/science.1139597](https://doi.org/10.1126/science.1139597)
- Wong, J. (2014). The Effects of High and Low Variability Phonetic Training on the Perception and Production of English Vowels /e/-/æ/ by Cantonese ESL Learners with High and Low L2 Proficiency Levels. *Proceedings of the 15th Annual Conference of the International Speech Communication Association, 524-528.* Retrieved from [https://repository.hkbu.edu.hk/hkbu\\_staff\\_publication/6234](https://repository.hkbu.edu.hk/hkbu_staff_publication/6234).
- Wong, J. W. S. (2012). Training the Perception and Production of English /e/ and /æ/ of Cantonese ESL Learners: A Comparison of Low vs. High Variability Phonetic Training. In *14th Australasian International Conference on Speech Science and Technology* (pp. 37–40). Sydney, Australia. Retrieved from <https://pdfs.semanticscholar.org/4801/008dea208d4474c108157bedbd672ab2202c.pdf>

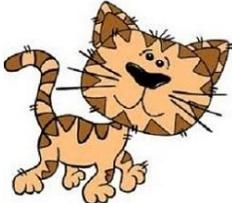
- Wong, J. W. S. (2014). The Effects of High and Low Variability Phonetic Training on the Perception and Production of English Vowels / e / - / æ / by Cantonese ESL Learners with High and Low L2 Proficiency Levels. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association* (pp. 524–528). Retrieved from [https://repository.hkbu.edu.hk/cgi/viewcontent.cgi?referer=https://scholar.google.co.uk/&httpsredir=1&article=7252&context=hkbu\\_staff\\_publication](https://repository.hkbu.edu.hk/cgi/viewcontent.cgi?referer=https://scholar.google.co.uk/&httpsredir=1&article=7252&context=hkbu_staff_publication)
- Wong, P. C., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, 28(4), 565-585. DOI: <https://doi.org/10.1017/S0142716407070312>
- Wong, P. C., Perrachione, T. K., & Parrish, T. B. (2007). Neural characteristics of successful and less successful speech and word learning in adults. *Human brain mapping*, 28(10), 995-1006. <https://doi.org/10.1002/hbm.20330>
- Wong, P. C., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature neuroscience*, 10(4), 420. <https://doi.org/10.1038/nn1872>
- 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>
- Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3), 37-52.
- 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 speaker. *The Journal of the Acoustical Society of America*, 143, 2013-2031. <https://doi.org/10.1121/1.5027410>
- Yang, J., Zhang, Y.L., Li, A., & Xu, L. (2017). On the Duration of Mandarin Tones. *INTERSPEECH*.
- Yip, M. (2002). *Tone. Cambridge textbooks in linguistics*. Cambridge: Cambridge University Press.
- Yokoyama, K., Jennings, R., Ackles, P., Hood, P., & Boller, F. (1987). Lack of heart rate changes during an attention-demanding task after right hemisphere lesions. *Neurology*, 37(4), 624-624. DOI: <https://doi.org/10.1212/WNL.37.4.624>
- Zare, P., & Riasati, M. J. (2012). The relationship between language learning anxiety, self-esteem, and academic level among Iranian EFL learners. *Pertanika Journal of Social Sciences and Humanities*, 20(1), 219-225.
- 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>
- Zhang, K., Peng, G., Li, Y., Minett, J. W., & Wang, W. S. (2018). The effect of speech variability on tonal language speakers' second language lexical tone learning. *Frontiers in psychology*, 9, 1982. <https://doi.org/10.3389/fpsyg.2018.01982>
- Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: a psycholinguistic grain size theory. *Psychological bulletin*, 131(1), 3. DOI: [10.1037/0033-2909.131.1.3](https://doi.org/10.1037/0033-2909.131.1.3)

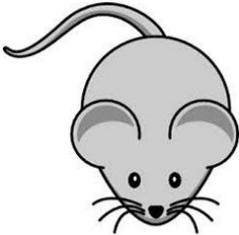
Zou, T., Chen, Y., & Caspers, J. (2017). The developmental trajectories of attention distribution and segment-tone integration in Dutch learners of Mandarin tones. *Bilingualism: Language and Cognition*, 20(5), 1017-1029. DOI: <https://doi.org/10.1017/S1366728916000791>

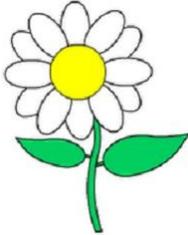
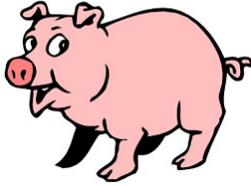
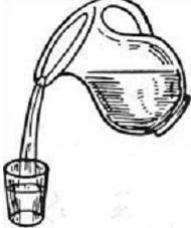
## Appendix A

Thirty-six pairs of Mandarin words and their corresponding pictures used in Study 1 and Study 2.

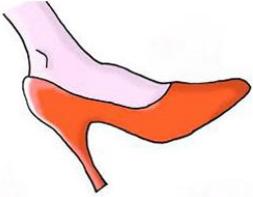
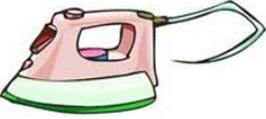
### *Trained stimuli:*

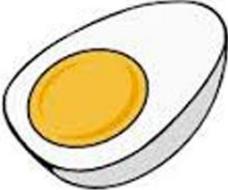
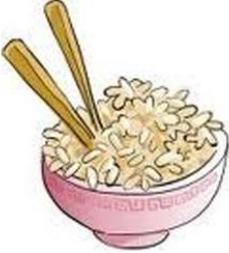
TONE 1	TONE 2
<p data-bbox="384 651 603 685">Chuāng (window)</p> 	<p data-bbox="1018 651 1182 685">Chuáng (bed)</p> 
<p data-bbox="435 985 552 1019">Māo (cat)</p> 	<p data-bbox="1018 985 1182 1019">Máo (anchor)</p> 
<p data-bbox="416 1299 571 1332">Qiān (swing)</p> 	<p data-bbox="1018 1299 1182 1332">Qián (money)</p> 

TONE 1	TONE 3 (ˇ)
<p data-bbox="427 253 560 286">Jiāo (glue)</p> 	<p data-bbox="1034 253 1166 286">Jiǎo (foot)</p> 
<p data-bbox="427 582 560 616">Shū (comb)</p> 	<p data-bbox="1034 582 1166 616">Shǔ (mouse)</p> 
<p data-bbox="427 911 560 945">Xuē (boot)</p> 	<p data-bbox="1034 911 1166 945">Xuě (snow)</p> 

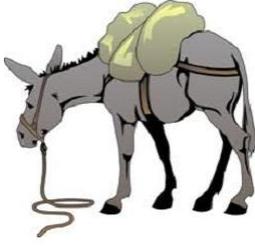
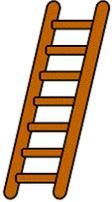
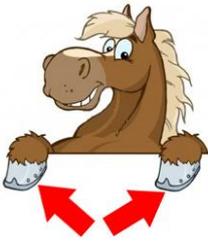
TONE 1	TONE 4
<p data-bbox="432 253 555 286">Bā (eight)</p> 	<p data-bbox="1031 253 1153 286">Bà (father)</p> 
<p data-bbox="416 560 571 593">Huā (flower)</p> 	<p data-bbox="1031 560 1153 593">Huà (paint)</p> 
<p data-bbox="432 889 555 922">Zhū (pig)</p> 	<p data-bbox="1031 889 1153 922">Zhù (pour)</p> 

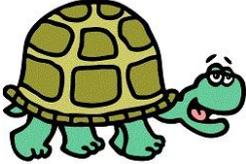
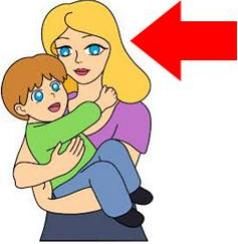
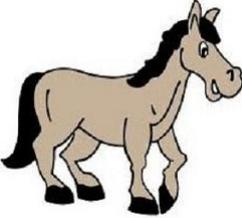
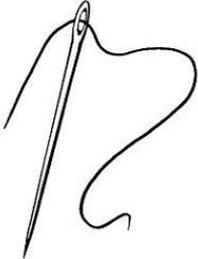
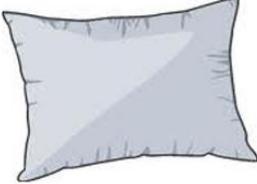
TONE 2	TONE 3 (ˇ)
<p data-bbox="437 253 552 286">Bí (nose)</p> 	<p data-bbox="1046 253 1161 286">Bǐ (pen)</p> 
<p data-bbox="411 573 577 607">Wán (to play)</p> 	<p data-bbox="1027 573 1174 607">Wǎn (bowl)</p> 
<p data-bbox="434 909 555 943">Niú (cow)</p> 	<p data-bbox="1024 909 1177 943">Niǚ (button)</p> 

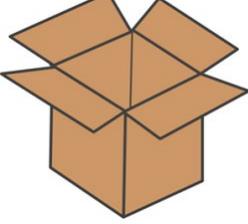
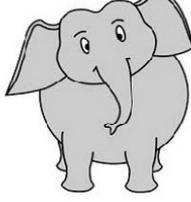
TONE 2	TONE 4
<p data-bbox="437 253 552 286">Dí (siren)</p> 	<p data-bbox="1038 253 1153 286">Dì (earth)</p> 
<p data-bbox="432 573 557 607">Xié (shoe)</p> 	<p data-bbox="1038 573 1153 607">Xiè (crab)</p> 
<p data-bbox="424 929 564 963">Yún (cloud)</p> 	<p data-bbox="1038 929 1153 963">Yùn (iron)</p> 

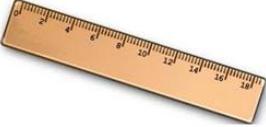
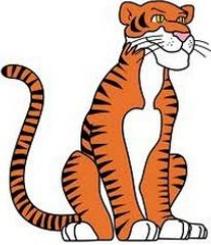
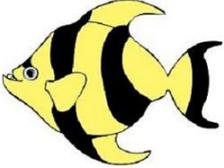
TONE 3 (ˇ)	TONE 4
<p data-bbox="389 255 600 286">Dǎn (brush, verb)</p> 	<p data-bbox="1038 255 1153 286">Dàn (egg)</p> 
<p data-bbox="440 611 549 642">Mǐ (rice)</p> 	<p data-bbox="1031 611 1161 642">Mì (honey)</p> 
<p data-bbox="437 967 552 999">Yǎn (eye)</p> 	<p data-bbox="1027 967 1165 999">Yàn (flame)</p> 

*Untrained stimuli*

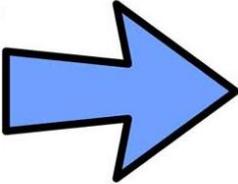
TONE 1	TONE 2
<p data-bbox="416 369 572 405">Shī (teacher)</p> 	<p data-bbox="1043 369 1150 405">Shí (ten)</p> <p data-bbox="1011 472 1182 600"><b>10</b></p>
<p data-bbox="427 676 561 712">Tuō (mop)</p> 	<p data-bbox="1011 676 1182 712">Tuó (to carry)</p> 
<p data-bbox="427 1008 561 1043">Tī (ladder)</p> 	<p data-bbox="1043 1008 1150 1043">Tí (hoof)</p> 

TONE 1	TONE 3 (ˇ)
<p data-bbox="411 257 576 291">Guī (tortoise)</p> 	<p data-bbox="1027 257 1166 291">Guǐ (ghost)</p> 
<p data-bbox="416 598 571 631">Mā (mother)</p> 	<p data-bbox="1027 598 1166 631">Mǎ (horse)</p> 
<p data-bbox="411 929 576 963">Zhēn (needle)</p> 	<p data-bbox="1015 929 1179 963">Zhěn (pillow)</p> 

TONE 1	TONE 4
<p data-bbox="416 253 571 286">Dēng (lamp)</p> 	<p data-bbox="1018 253 1173 286">Dèng (bench)</p> 
<p data-bbox="443 649 544 683">Kū (cry)</p> 	<p data-bbox="1018 649 1173 683">Kù (trousers)</p> 
<p data-bbox="427 981 560 1014">Xīāng (box)</p> 	<p data-bbox="1002 981 1189 1014">Xiàng (elephant)</p> 

TONE 2	TONE 3 (ˇ)
<p data-bbox="424 253 564 286">Chí (spoon)</p> 	<p data-bbox="1034 253 1161 286">Chǐ (ruler)</p> 
<p data-bbox="397 616 592 649">Hú (moustache)</p> 	<p data-bbox="1038 616 1158 649">Hǔ (tiger)</p> 
<p data-bbox="440 956 549 990">Yú (fish)</p> 	<p data-bbox="1026 956 1171 990">Yǔ (feather)</p> 

TONE 2	TONE 4
<p>Mó (mushroom)</p> 	<p>Mò (mill)</p> 
<p>Shé (snake)</p> 	<p>Shè (house)</p> 
<p>Wá (baby)</p> 	<p>Wà (sock)</p> 

TONE 3 (ˇ)	TONE 4
<p data-bbox="411 257 580 293">Bǎo (treasure)</p> 	<p data-bbox="995 257 1203 293">Bào (newspaper)</p> 
<p data-bbox="421 573 571 609">Dǎo (island)</p> 	<p data-bbox="1034 573 1165 609">Dào (road)</p> 
<p data-bbox="411 929 580 965">Jiǎn (scissors)</p> 	<p data-bbox="1027 929 1171 965">Jiàn (arrow)</p> 

## Appendix B

### Word set used in Study 3, Training and Pinyin Reading test

Note that all the base syllables in this list are accompanied by all four Mandarin tones and all of them are genuine Mandarin words.

For Pinyin reading (see Section 4.3.3.3): Words marked in green shares the same (approximate) phonology in Mandarin and English thus may be correctly pronounced by English speakers without training. Words marked in yellow share some phonological segments. When estimating baseline performance for the Pinyin Measure (i.e. what would be expected for learners with no knowledge of Mandarin) for the BF calculations I assumed that: the words in the green cells would always be pronounced correctly, the words in the yellow cells would have a 50% chance of being pronounced correctly and the words in the clear cells would never be pronounced correctly. This led to an estimation of 8/18 as the baseline score.

Word	Combination with Tone			
	Tone 1	Tone 2	Tone 3	Tone 4
shì	Shī	Shí	Shǐ	Shì
chuang	Chuāng	Chuáng	Chuǎng	Chuàng
fang	Fāng	Fáng	Fǎng	Fàng
shu	Shū	Shú	Shǔ	Shù
xue	Xuē	Xué	Xuě	Xuè
ma	Mā	Má	Mǎ	Mà
ba	Bā	Bá	Bǎ	Bà
xiang	Xiāng	Xiáng	Xiǎng	Xiàng
ku	Kū	Kú	Kǔ	Kù
yu	Yū	Yú	Yǔ	Yù
niu	Niū	Niú	Niǔ	Niù
wan	Wān	Wán	Wǎn	Wàn
mo	Mō	Mó	Mǒ	Mò
yun	Yūn	Yún	Yǔn	Yùn
xie	Xiē	Xié	Xiě	Xiè
yan	Yān	Yán	Yǎn	Yàn
dao	Dāo	Dáo	Dǎo	Dào
mi	Mī	Mí	Mǐ	Mì

## Appendix C

Word set used in Study 3, Four Interval Oddity Task

(Note that although English translations are provided, this study did not involve semantic learning)

List	Tone Contrast	Item 1	Item 2
1-2		Māo (cat)	Máo (anchor)
		Tuō (mop)	Tuó (to carry)
		Qiān (swing)	Qián (money)
1-3		Jiāo (glue)	Jiǎo (foot)
		Guī (tortoise)	Guǐ (ghost)
		Zhēn (needle)	Zhěn (pillow)
1-4		Dēng (lamp)	Dèng (bench)
		Huā (flower)	Huà (paint)
		Zhū (pig)	Zhù (pour)
2-3		Bí (nose)	Bǐ (pen)
		Chí (spoon)	Chǐ (ruler)
		Hú (moustache)	Hǔ (tiger)
2-4		Dí (siren)	Dì (earth)
		Shé (snake)	Shè (house)
		Wá (baby)	Wà (sock)
3-4		Dǎn (brush, verb)	Dàn (egg)
		Bǎo (treasure)	Bào (newspaper)
		Jiǎn (scissors)	Jiàn (arrow)