Training Child Learners On Non-Native Vowel Contrasts with Phonetic Training: The Role of Task and Variability

Gwen Brekelmans,[1,2] Bronwen G. Evans,[3] Elizabeth Wonnacott [2,4]

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

[2] *Department of Language and Cognition, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom*

[3] *Department of Speech, Hearing & Phonetic Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom*

[4] *Department of Education, University of Oxford, 15 Norham Gardens, Oxford OX2 6PY, United Kingdom*

**Running title:** Variability in phonetic training for children

Corresponding author: Gwen Brekelmans, g.brekelmans@qmul.ac.uk

[This preprinted manuscript is currently in press at Language Learning].
[Last date updated: July 2024]

# ABSTRACT

Substantial research suggests high variability (multi-talker) phonetic training helps second language adults improve differentiation of challenging non-native speech sounds. Is such training also useful for L2 children? Existing studies have mixed findings and important limitations. We investigate the potential benefits of computerised phonetic training for 50 Dutch 7-year-olds and 39 11-year-olds trained on English vowel contrasts in a two-week study, in a classroom setting. Half received multi-talker, half single-talker input (HV versus LV) with learning evaluated by a battery of tests.  Both groups improved in training, however, 11-year-olds improved more. Moreover, 11-year-olds showed generalisation to novel talkers while 7-year-olds did not, with Bayes Factor analyses providing evidence for the null. Generalisation in 11-year-olds was no greater following HV than LV input, with evidence for the null on one of two tasks where generalisation was found. Results are discussed in terms of the interplay between age, task demands and talker variability.

## INTRODUCTION

Acquiring non-native speech sounds can be challenging. Seminal papers by Logan, Lively and colleagues in the 1990s established the benefits of High Variability Phonetic Training (HVPT) – comprising targeted practice in distinguishing minimal pairs, with feedback, using varied tokens from multiple talkers. This has become a standard methodology in the field (cf. Brekelmans et al., 2022), though the specific benefits of high variability are unclear. Moreover, these studies have typically focussed on adult learners; less work has explored whether children benefit from phonetic training, and whether variability is beneficial for their learning. A small number of studies have investigated the efficacy of HVPT for children, but findings are mixed and the studies themselves have limitations: any potential learning is evaluated using a small number of tasks with training focussing on only a single contrast. The current study expands on this literature by testing the use of computerised phonetic training in a classroom-context, with two groups of primary school age children: 7- and 11-year-old Dutch learners of English. We compare multiple-talker (HV) versus single-talker (LV) input and test learning and generalisation using a wider range of tasks, and train on several vowel contrasts.

### Phonetic training in adults

In formative work, Logan et al. (1991) trained Japanese learners of English on /l/-/r/, using natural rather than synthesised speech to exemplify the contrast in multiple phonetic contexts with items spoken by multiple talkers. Participants were tested on their ability to identify /r/ and /l/ both pre- and post-training: they showed improvement for trained items and, critically, generalisation to novel voices and items. In a follow-up experiment, Lively et al. (1993) demonstrated that generalisation to new talkers was dependent on training with multiple talkers, i.e. having high variability training input. This benefit of high variability input for generalisation is theoretically and intuitively appealing: encountering variability helps the listener recognise which acoustic cues are irrelevant and focus on key cues to identifying the phonemes. Since then, phonetic training – generally incorporating talker variability - has been shown to be highly effective in improving adult learners' perception of many different types of contrast (e.g. *vowels*: Fuhrmeister & Myers, 2020; *consonants*: Sadakata & McQueen, 2013; *lexical tones*: Antoniou & Wong, 2015). These effects have been shown to hold over time (e.g. Bradlow et al., 1999) and in some studies, to transfer to production (e.g. Bradlow et al., 1997).

Perhaps suprisingly, the *specific* advantage of using *high* (HV) rather than low variability (LV) input is less well-established: we only found it reported in three additional studies. Wong (2012) trained Cantonese learners on English vowel contrasts, and found that learners in both HV and LV conditions improved after training. However, those in the HV condition outperformed those in the LV condition and showed greater generalisation to novel talkers and items. Sadakata & McQueen (2013) trained Dutch adults on a Japanese geminate contrast, varying talker and context across training. Change from pre- to post-test in an identification task indicated more learning and generalisation to new items containing the same contrasts and to new talkers after HV training than LV training, although there was no such difference between HV and LV training in a discrimination task. Other phonetic training studies, however, have failed to find an overall high variability advantage: Perrachione et al. (2011) and Sadataka & McQueen (2014) found HV input detrimental for learners of lower perceptual aptitude, while others (e.g., Deng et al., 2019; Wiener et al., 2020) found no variability differences. Further, a recent large-scale replication of the Logan et al. (1991) study showed that participants showed substantial learning and generalisation to novel talkers and items after both HV and LV training, but no evidence for a benefit of multiple talkers on generalisation (Brekelmans et al., 2022).

Given that there are clear theoretical reasons to expect that high variability input aids generalisation to novel talkers and items (see e.g. Kleinschmidt & Jaeger, 2015), it is unclear why previous studies do not consistently find better generalisation for high variability input. One potential explanation is that there are methodological differences between the studies, e.g., in terms of what tasks are used and which phonetic contrasts are trained. For example, some studies only train a single vowel contrast that is known to be difficult (e.g. Logan et al., 1991;

Giannakopoulou et al., 2013), while others train a combination of vowels and consonants (e.g., Cebrian & Carlet, 2014). Still others have shown that the number of contrasts affects learning, with training on the full vowel system more successful than training on a smaller number of difficult contrasts (Nishi & Kewley-Port 2007, 2008). It would perhaps not be surprising if the number and difficulty of the phonetic contrasts being trained might place different demands on the learner, which could potentially interact with the benefit of variability. In terms of task type, Sadakata & McQueen (2013) found stronger generalisation following high rather than low variability phonetic training when training was completed using an identification task, but no such difference with a discrimination task.

Another possibility raised in the literature is that benefits of high variability may be masked by interactions with learner aptitude. Support for this comes from two studies – Perrachione et al. (2011) and Sadakata & McQueen (2014) – which found that when applying phonetic training to lexical tones, HV input was detrimental for generalisation to novel talkers in learners who had weaker perceptual abilities. This could be due to a trade-off between the benefits of variability, and the greater initial processing cost of adapting to multi-talker input. This burden could be greater in learners with weaker perceptual abilities, offsetting any variability benefit for generalisation. This would be in line with literature showing that adapting to multiple talkers in L1 speech processing is initially detrimental to word recall and processing (Martin et al., 1989; Mullennix et al., 1989). It also fits with evidence from HVPT studies which track performance *during* training and which often find lower performance on the HV than LV materials (Perrachione et al., 2011; Dong et al., 2019; Giannakopoulou et al., 2017; Evans & Martín-Alvarez, 2016). This suggests that although variability may ultimately benefit generalisation, this is balanced against the immediate greater difficulty of processing more varied materials.

While it seems intuitive that some learners might have more difficulty with HV input, other studies have failed to find an effect of individual aptitude on L2 learning. In a study investigating acquisition of Mandarin tone by native English speakers, Dong et al. (2019) not only found no overall variability benefit, but also no evidence of an interaction with individual aptitude. Similarly, Wong (2014) found no modulating effect of learner proficiency on acquisition of English /e/-/æ/ by native Cantonese learners, although *both* high and low proficiency groups (rather than neither) seemed to benefit from variability. In sum, while a substantial literature shows that phonetic training is successful in adults, a specific benefit of increased variability is not consistently found.

## Phonetic training with child learners

On average 80% of children in EU primary education learn a foreign language in school (Eurostat, 2019). Computerised training has clear potential for use in language classrooms, particularly in contexts where native-speaker input may be lacking. A handful of studies has investigated HVPT in children, generally focused on investigating whether younger learners show faster or more efficient learning than older learners (based on the expectation of plasticity effects, e.g. Flege & MacKay, 2004). Results are mixed. Some studies find similar amounts of improvement in perceptual performance across adults and children (e.g. Wang & Kuhl, 2003; Giannakopoulou et al., 2017; Heeren & Schouten, 2010). Others *have* found benefits for younger learners: Giannakopoulou et al., (2010; 2013) found greater pre- to post-improvement in Greek-speaking 7-8-year-olds than in adults trained on English /iː/-/ɪ/, and Shinohara and Iverson (2021) found that Japanese speaking 8-12-year-olds and 15-18-year-olds improved more than adults when trained on the English /r/-/l/ contrast. However, inconsistent with the younger-is-better claim, this study also found that a younger group of 6-8-year-olds did not improve at all. The authors suggest that this may be due to the fact that responses were made using orthographic symbols, and that these younger children may have been less familiar with the concept that letters can consistently map to sounds.

An additional question is whether HV input is specifically beneficial. Since natural language input is rife with variability across different dimensions, and in naturalistic settings children tend to be better phonetic learners than adults (e.g., Flege & MacKay, 2004), one could imagine younger learners might be particularly adept at dealing with HV input and might benefit from this in learning. However, to the best of our knowledge, to date, only two studies have directly compared HV/LV phonetic training in children, with contradictory results. Evans & Martín-Alvarez (2016) trained 44 Spanish 9-12-year-olds on the English /iː/-/ɪ/ contrast in either an HV (four talkers) or LV (one talker) condition using a picture identification task and testing potential improvements in discrimination (words and non-words, new talkers) and transfer to production. Only children in the LV condition improved in production, and although there was an HV benefit for vowel discrimination this was limited to novel items (i.e., non-words). Giannakopoulou et al. (2017) trained Greek-speaking 8-year-olds using the same stimuli and similar methods (though not testing production) and did *not* find an HV benefit in the discrimination test. Instead, they showed greater generalisation in the LV condition. They suggested that this LV benefit could be due to greater difficulty in processing multi-talker training materials (cf. Perrachione et al., 2011; Sadakata & McQueen, 2014. Martin et al., 1989; Mullennix et al., 1989).

Interestingly, a parallel literature on vocabulary training suggests that adults and children might differ in whether they benefit from variability. Here, there is substantial evidence that adults benefit from learning words from multiple talkers compared with a single talker (cf. Barcroft & Sommers, 2005; Sommers & Barcroft, 2011). However, word learning studies with children have *not* found this advantage (Crespo & Kaushanskaya, 2019; Crespo, Vlach and Kaushanskaya, 2023; Sinkevičiūtė, Brown, Brekelmans & Wonnacott, 2019). For example, Sinkevičiūtė et al. (2019) compared 7-year-olds, 11-years-old, and adults in a novel word-learning experiment. While adults showed the expected HV benefit, this was not true of either group of children. Indeed, 7-8-year-olds showed a *low* variability advantage during training itself, suggesting an impact of the difficulty of processing more variable input. The authors discuss this in terms of a trade-off between potential benefits of variability for generalisation and processing difficulties, which might be greater in younger learners due to cognitive differences such as lower working memory capacity (e.g. Alloway et al., 2006).

In sum, there is evidence that children can benefit from phonetic training, but the effects of age and variability are unclear. A final limitation of all of the studies conducted with children, is that in every case they trained a single vowel contrast. This is despite the fact that in adults, there some evidence that training multiple contrasts is helpful (Nishi & Kewley-Port, 2007), and it clearly limits the utility of the training paradigm in educational contexts.

## Difficulties in interpreting null results

It is clear from the above that there are mixed results regarding the benefits of variability in training, and interactions with age and task type. An additional difficulty in interpreting these results is that where studies do present null results, the type of statistics used for inference (p-values) mean that it is impossible to draw conclusions about the original hypotheses of p-values in the literature ($p>.05$ does not tell us that we should increase our confidence in the null hypothesis, despite common misinterpretation). Just two of the studies cited above used statistical methods that allow for providing evidence against a hypothesis: Dong et al. (2020), who found evidence against the hypothesis that there is a HV benefit in tone learning, and Sinkevičiūtė et al. (2019) who found evidence against a HV benefit in 7-year-old word learning. Critically, these studies used Bayes Factors for statistical inference, which *can* distinguish whether there is evidence for the null (i.e. against the predicted hypothesis), or whether the evidence is ambiguous and the data are insensitive (Dienes, 2014). We take this approach in the current work.

## The current study

Overall, the literature reviewed above suggests that while there is substantial evidence phonetic training can boost learning and generalisation of L2 phonetic contrasts in adults, this is less well-studied in children. In addition, despite clear theoretical reasons to expect that generalisation over acoustic features will be boosted by talker variability, this effect in phonetic training is not well-established, with some studies suggesting that effects may be modulated by task type and age. A parallel literature in vocabulary learning has found evidence for a related benefit of variability in adults, but not in children. Moreover, there is a difficulty in interpreting the null results reported in many of the papers.

The goal of the current study is to further explore whether phonetic training may be effective in an intervention with primary-school-aged children. We recruited Dutch learners of English aged 7-8-years and 11-12-years. To establish educational relevance, we embedded the study in a classroom-context and used a word learning task requiring learners to differentiate multiple challenging vowel contrasts. The study comprised three stages: pre-test, training, and post-test. We tested whether there are specific benefits of HV over LV input, and differences between older (11-year-olds) versus younger (7-year-olds) learners. Note that this choice of age-groups allowed us to recruit children from within the same school (11-12-year-olds are the oldest age group in Dutch primary schools) and still look for age differences. We acknowledge that age and years of English experience are naturally confounded in this context, so our analyses with respect to age focus on pre- to post- test improvement. Half of the participants were trained on multi-talker (HV) and half on single-talker (LV) input. To avoid potential difficulties with orthography (cf. Shinohara & Iverson, 2021), children were trained using an identification task in which they heard words that were displayed on the screen both in their written form and using a picture. Performance during training was tracked and generalisation over the training input was assessed using a battery of identical pre- and post-tests. These tasks were designed in light of the literature reviewed above and consisted of a Discrimination, Picture Identification, Vocabulary and Orthography Identification task. Note that both phoneme discrimination and identification were tested, since differences between these are reported (e.g. Evans & Martín-Alvarez, 2016, Sadakata & McQueen, 2013; Heeren & Schouten, 2010). There was a vocabulary test given the clear educational relevance and in light of the literature on variability benefits in this area (e.g. Sinkevičiūtė et al., 2019). Separate identification tests used orthography versus pictures, bearing in mind potential challenges with orthography for younger learners (cf. Shinohara & Iverson, 2021).

We take the approach of directly testing a set of precise hypotheses based on the literature reviewed above – see table 1 – which were pre-registered (https://doi.org/10.17605/OSF.IO/KMSPQ)[2] and tested using Bayes Factors. Note that in some cases we have two opposing hypotheses, listed as a) and b) [3] these were pre-registered as separate hypotheses, and tested with two distinct tests (with different pre-registered priors for Bayes Factor analyses). Where appropriate, we also tested whether overall performance was above chance [4].

| Prediction | Supporting literature |
|---|---|
| 1) Improvement in both age groups | Giannakopoulou et al., 2017; Evans & Martín-Alvares, 2016; Shinohara, 2014; Wang & Kuhl, 2003; Giannakopoulou et al., 2010, 2013; Heeren & Schouten, 2010 |
| 2a) older children will improve more than younger children<br><br>2b) younger children will improve more than older children | a) Shinohara, 2014 (comparison 6-7-year-olds versus 8-18-year-olds); Shinohara & Iverson, 2021 (comparison 6-8-year-olds versus 8-12-year-olds & 15-18-year-olds)<br><br>b) Giannakopoulou, 2013 (comparison 7-8-year-olds versus adults) |
| 3) There will be better performance during training in the LV group. | Giannakopoulou et al., 2017; Evans & Martín-Alvarez, 2016 (training data with adults and 7-8-year-olds) |

| | |
|---|---|
| 4a) Generalisation to novel talkers and items (i.e., improvement from pre- to post-test ) will be greater in the HV condition.<br><br>4b) Generalisation to novel talkers and items (i.e., improvement from pre- to post-test) will be greater in the LV condition. | Barcroft & Sommers, 2005 (vocabulary); Lively et al., 1993<br><br>Giannakopoulou et al., 2017 (result in discrimination test with 7-year-olds) |
| 5) Any benefit of HV in generalisation will be stronger in older than younger children | Giannakopoulou et al., 2017 |
| 6) Any generalisation benefit of HV will be stronger in novel items | Evans & Martín-Alvarez, 2016 |

Table 1. Pre-registered hypotheses and supporting literature.

## METHODS

### Participants

Participants were recruited from *Groep 4* (7-8 years old), and *Groep 8* (11-12-years old) in a school in the south of the Netherlands. Opt-out parental consent was used; at any point in the study children could decide not to participate and parents/carers could withdraw them. Each child participated in 10 sessions, each on separate days over a maximum of 12 consecutive school days, receiving daily stickers and a certificate on completion. Our preregistered target sample was 48 7-8 year-olds and 48 11-12 year-olds. In fact, due to whole classes wishing to participate 51 11-12 year-olds and 58 7-8 year-olds were tested. Following the preregistered exclusion criteria, data from 20 of these were excluded due to: not completing at least 60% of training; having dyslexia/hearing impairments/other language disorders; having English-speaking parents/carers; having lived in an English-speaking country for more than two months. This left

50 7-8 year-olds (mean age = 7;6, SD = 4 months, range 6;10–8;3 years) and 39 11-12 year-olds (mean age = 11;7, SD = 4 months, range 10;9–12;0 years). Additionally, for each task, participants were excluded from analyses if either their pre-test or post-test data was missing. Full details of participants' language background and participant numbers included in each task can be found on OSF.

The age groups differed slightly in previous English instruction in school: 7-8 year-olds had not yet started English, while 11-12 year-olds had received one English lesson per week for about a school year (i.e. 40 weeks). These English lessons focussed on communicative content rather than pronunciation, and were taught by native Dutch speakers with no specific pronunciation training. Children reported low English exposure outside of school, and while 11-12 year-olds had a little more exposure than 7-8 year-olds, this was matched for the variability conditions within the groups (see Supplementary information A2 on OSF).

**Stimuli**

Stimuli were recorded by 7 native speakers of Standard Southern British English (5 female, 2 male) in a sound-attenuated booth at a sampling rate of 44.1 kHZ 16-bit samples per second. They were later downsampled to 22.05kHz and band-pass filtered from 60-20,000 Hz with a smoothing factor of 10. Finally, intensity was scaled to 70 dB. All processing was carried out in Praat (Boersma & Weenink, 2015). Stimuli were presented using ProRec (Huckvale, 2016). All stimuli are available on OSF at https://doi.org/10.17605/OSF.IO/BGDXP.

***Trained stimuli***

Trained stimuli consisted of 16 monosyllabic CVC minimal pairs,differing in the vowel. Consonants were chosen which are present in the consonant inventory of both Dutch and English. There were 3 vowel contrasts which were expected to be difficult (Collins & Mees, 2003): GOOSE-FOOT /uː/-/ʊ/, STRUT-LOT /ʌ/-/ɒ/, DRESS-TRAP /e/-/æ/, and a control contrast FLEECE-THOUGHT /iː/-/ɔː/ which should be easy for Dutch learners as it has a close native counterpart. The latter provides a check on whether participants understood the task, and to boost confidence/task appeal by having easier items.

Each vowel contrast was exemplified by 4 minimal pair items (8 words), making 32 trained items (16 trained pairs; full stimuli list in Supplementary information A3 on OSF). Words were chosen for imageability and, as far as possible, transparency and consistency of orthography (since pictures and written words were used in training). All words were deemed appropriate for and famliar to primary school children,as established in initial pilot tests. The trained stimuli were used in both Training and all pre/post-test tasks.

***Novel stimuli***

The Discrimination and Orthography Identification tasks also included novel items to test generalisation (see Supplementary information A3 on OSF): 32 novel items (16 pairs), were used in Discrimination task (making 64 items in total: 32 trained, 32 novel); A 32 different novel items (16 pairs), were used in the Orthography taks (again making 64 items in total: 32 trained, 32 novel). In the latter, since English orthography is known to be opaque, the novel pseudoword items were created to be as transparent as possible keeping vowel orthography identical to that used in the trained items.

**Design**

The study had a pre-/post-test design, with data also collected during training. Talker variability was manipulated between-participants. Talkers were drawn from a total of seven native speakers of Standard Southern British English; assignment of talkers was counterbalanced. In HV training, participants heard four different speakers (two male, two female) while in LV training stimuli were spoken by a single talker (one of the four used in HV, with voices rotated across versions to ensure all talkers were equally represented). In the HV condition, stimuli from the four speakers were presented in separate blocks, as trial-by-trial speaker alteration has been shown to be detrimental to training performance in phonetic training paradigms (e.g. Perrachione et al., 2011). Speaker order was counterbalanced across blocks. Pre- and post-tests were identical across conditions, and – in order to test generalisation to novel voices – all involved novel speakers (three females, rotated between tasks) who had not been encountered during Training. Full counterbalancing details are in Supplementary information A4 on OSF.

**Procedure**

Testing took place in groups in the children's classroom, and was administered by the first author. Children completed the tasks individually, each on a laptop running Gorilla (www.gorilla.sc, Anwyl-Irvine et al., 2019) and using their own in-ear headphones. They completed the pre-test the day before their first training session. They then completed 8 training sessions on separate days, and the post-test the day after the last training session (ensuring any temporary phoneme categories built up during training were no longer maintained; Heeren & Schouten, 2010). Figure 1 gives an overview of the procedure. See Supplementary information A1 on OSF for task screenshots .

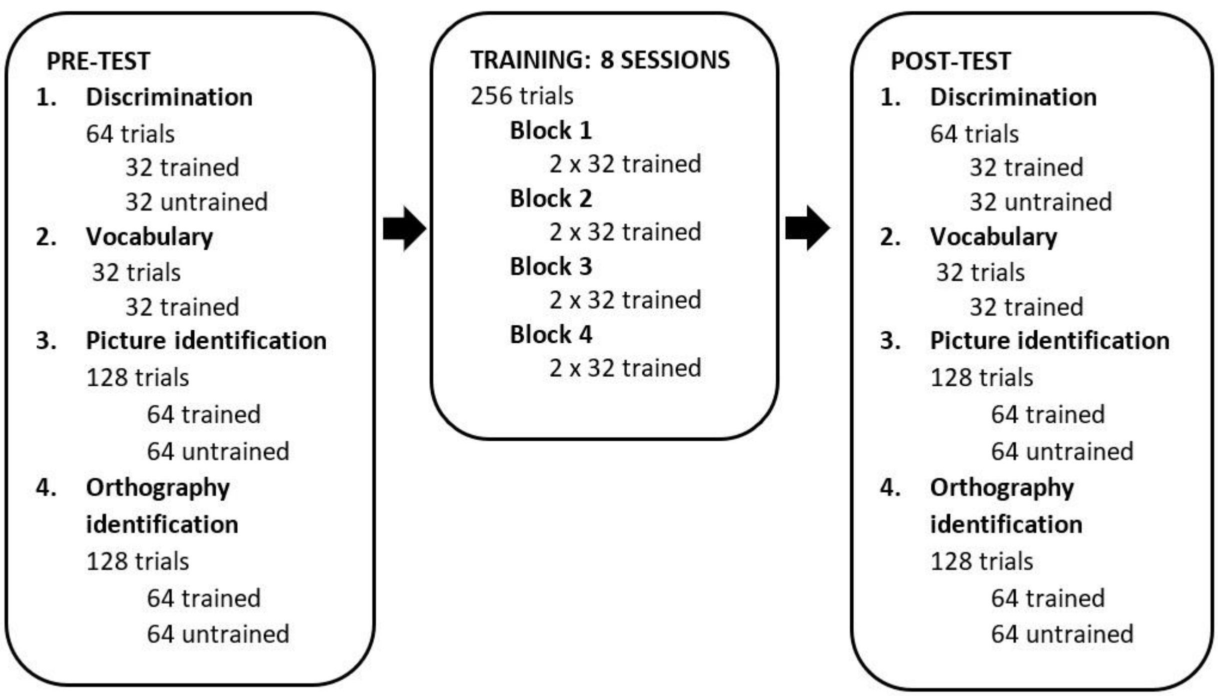


Figure 1. Overview of the Training and pre/post-tests procedure, including the number and type

of experimental trials per task. Task numbering shows the order in which tasks were conducted. Pre- and post-tests were administered in a separate session.

***Training***

The training task was a two-alternative forced-choice (2AFC) identification task. In each trial, participants heard a word and viewed two pictures, with accompanying written words, representing the target and foil (i.e., the minimal pair of the target). The task was to select the correct word. They then received feedback, a reward (an animated coin), and were shown the correct response. There were 8 training sessions, 4 blocks each, with each block comprising 64 trials – twice with each of the 32 training words as target – spoken by a single talker. In total, 256 trials were presented per session. Trial order was randomised within blocks on a participant-by-participant basis.

***Pre/post-tests***

*Discrimination task*

This was a three-alternative forced-choice (3AFC) oddity task. In each trial, three animated figures appeared on the screen; a speech bubble appeared when the corresponding word was played. Participants heard three words (ISI 200ms), two of which were different tokens of the same word (e.g., *bed, bed*), and one of which was a minimal pair foil containing the other vowel from the trained contrast (e.g., *bad*). After all words had played, participants indicated which was the odd one out by clicking on the corresponding speech bubble. Each token was spoken by a different talker.

There were 16 trained and 16 untrained minimal pairs. Each pair occurred once with each member of the pair as target (64 trials in total). Trial order was pseudo-randomised. No feedback was provided.

*2AFC Picture identification task*

This was identical to the training task except that (1) only pictures but no written words were used (2) there was no feedback. There were 128 trials: 32 minimal pairs used in training each occurring once as the target; 32 non-minimal pair combinations that differed in more than just the vowel, each occurring once as the target. Presentation of minimal pairs and non-minimal pairs was blocked, with the order randomised within blocks. Note that (1) it is not possible to include novel *items* in this test, since it depends on knowledge of picture/item correspondences, and (2) including the non-minimal pair items provided an additional test of *vocabulary knowledge* without requiring an ability to discriminate the vowels. No feedback was provided.

*2AFC Orthography identification task*

This was like the training task except that (1) only written words but no pictures were used (2) there was no feedback. The task comprised 128 trials: 32 minimal pairs from training, each occurring once as the target; and 32 novel minimal pairs each occurring once as the target. Presentation of trained and novel items was blocked, with a randomised order within blocks. No feedback was provided.

## Analyses and statistical approach

Data were analysed using both Frequentist and Bayesian statistics.

### ***Model building***

Data were analysed using generalised logistic mixed effects models (glmer) (Baayen et al., 2008) using the lme4 package (v1.1.33; Bates et al., 2015) for the R environment (v4.3.0; R Core Team, 2023). Data was visualised using ggplot2 (v3.4.2; Wickham, 2016) and ggpirate (v0.1.2; Braginsky, 2018).

Performance for each task was analysed in separate models. The dependent variable across all analyses was accuracy. The approach was to include all experimentally manipulated variables and their interactions for each task as fixed factors in the models, regardless of whether they contributed to the model. Key fixed effects were *variability* (HV/LV), *session* (1-8 for training, pre/post for all other tasks), and *item novelty* (in tasks where this was manipulated, i.e., trained/novel for Discrimination and Orthography Identification, minimal pair/non-pair for Picture Identification) and all the interactions between them. As preregistered, the two age groups were first analysed separately, followed by targeted comparisons including *age group* as a variable[5]. In total, analyses were run in 11 separate models.

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

A centred coding was used for all predicting variables. This means that the beta estimate for the intercept corresponds to the grand-mean and each fixed effect is computed as averaged over all levels of other factors (i.e. they are main effects rather than simple effects). All reported models converged using the Bound Optimization by Quadratic Approximation (BOBYQA optimization: Powell, 2009).

In addition to the experimental factors, factors to control for *vowel contrast* (DRESS-TRAP, GOOSE-FOOT, STRUT-LOT, and FLEECE-THOUGHT) and *talker* (4 different talker identities in training; 3 in the tests) were included in the analyses since these factors are expected to contribute variance to the data (different contrasts might be perceptually more or less difficult; talker idiosyncrasies contribute to difficulty). Both are included as fixed effects in the model and are coded as a set of centred contrasts (3 for vowel contrast[6]; for talker these were 4 for training and 3 for the tests) so that the intercept continues to represent the grand mean. Conceptually, *talker* is a random effect, but this cannot have fewer than 6 levels (Bolker, 2019) so is entered as a fixed effect. *Talker* was only included when it contributed to the model fit (with $p>.2$ similar to what is suggested for random effects by Matuschek et al., 2017). These factors were not interpreted nor included in interactions. Full model structures for each analysis are reported on OSF.

### ***Inference criteria – Frequentist and Bayesian analyses***

Although models included all experimentally manipulated variables and their interactions, we only interpret and report statistics for fixed effects relevant to our hypotheses. The key inferential statistic used to evaluate these hypotheses is the Bayes Factor (BF), which computes the strength of evidence for the hypothesis (H1) over the null hypothesis (H0), or vice versa. These differ from p-values, which do not provide the option to evaluate the strength for the H0. Since Bayes Factors are less familiar, p-values (automatically computed from z-values by the *lme4* package (Bates et al., 2015)) are additionally reported.

Bayes Factors were computed using the method advocated by Dienes (2008, 2015): Specifically, for each hypothesis, we compute the ratio of the likelihood of the relevant data under H1 to the likelihood under the null. For our model of the null we use a point null. For our model of H1 (also known as the prior) we use a distribution representing the plausibility of different sized effects. Specifically, we use a half normal with a mean of zero and a standard deviation set to a rough estimate of the size of effect we predict. The use of the half normal means that we are testing a directional, one-sided prediction (appropriate since in each case we are testing a directional hypothesis) and setting the mean at zero instantiates the assumption that smaller effects are more plausible[7]. We use an implementation of the Dienes 2008 calculator (see Dienes 2008, 2014) to compute the Bayes Factor for a given data set using these distributions (using R code provided by Baguely & Kayne, 2010). This method requires three numbers: (i) an estimate of a mean difference (between conditions, or between a mean and some baseline) (ii) the associated standard error; (iii) a rough estimate of the *predicted* mean difference under H1, *x.* The third value is used to inform the model of H1 (a half normal with mean=0 and sd=*x*). The first two form a sufficient summary of the data for the calculator. Throughout we take (i) from the estimate (beta) for the relevant fixed effect in the logistic mixed effect model and we take (ii) from its associated standard error. Note that these values are in log odds space, meeting the normality requirements of the calculator (for a recent paper using this method see Brekelmans et al., (2022)). The key issue is how to determine appropriate values of *x* to inform our model of H1. While we do not need precise values (since we model H1 as a distribution) the use of unprincipled estimates of predicted effect size (such as the defaults provided in some software for computing Bayes Factors) may lead to spurious evidence for H0 (Nicenboim & Vasishth, 2016). We attempted to determine estimates as follows: In some cases, we use estimates taken from equivalent analyses in Giannakopoulou et al. (2017), which used a similar task design. Where this was not possible/appropriate, values from within the current data set were used. Generally, this involved using estimates from independent participants, (e.g., using the sample estimate for the effect of variability in 7-8 year-old children as the predicted effect size for 11-12 year-olds, and vice versa). Where this was not possible, we determined a plausible approximate effect size *x* by first determining a likely plausible maximum and then setting *x* to *half* this value [since for a half normal with a mean of 0 and sd=x, 2x is a rough maximum; see Silvey et al., (2024) and Dienes & McLatchie, 2018)].

Our choice of values of H1 for each test was pre-registered (except where noted – see endote 8) and is laid out in Supplementary information A5 on OSF. BF results were reported using the notation $BF_{H(0,x)}$ to denote a Bayes Factor where *x* is the standard deviation of the half normal used to model H1 (following Dienes, see https://osf.io/hzcv6/).

Bayes factors were interpreted using the convention in Jeffreys (1961) and expanded by Dienes (2014), where BF>3 indicates substantial evidence for H1, a BF between 1/3 and 3 indicates ambiguous evidence (the data is insensitive to test the hypothesis), and BF<1/3 indicates substantial evidence for H0. Since there is subjectivity in the choice of value to inform H1, robustness regions are included as a type of sensitivity analysis, showing the range of estimates of H1 (i.e., the value used as the standard deviation of the half normal) for which our data would support the same conclusion (e.g., if the evidence for H1 is substantial with the criteria BF>3, the range of estimates which would also yield a BF>3). The larger the robustness region, the wider the range of values of H1 estimates that would support the same conclusion.

This shows that our conclusions are robust even if there are small but reasonable differences in the potential assumptions we used to inform the H1 estimates.

This robustness region is noted as [$x_1$, $x_2$] with $x_1$ being the smallest standard deviation and $x_2$ the largest standard deviation (as recommended by Dienes, 2017). To compute the ranges, values were tested for values from a smaller possible difference of 0 to a log-odds score corresponding to maximum plausible difference (between chance and ceiling performance of all but one trial correct. Note that there will always be some value which provides evidence for H0, where this was not found within the range tested, the end point of the tested range is noted as e.g., >4.8299). Robustness Regions should be interpreted bearing in mind that larger values of H1 bias evidence for the *null,* whereas smaller values bias in favour of H1.

## RESULTS

Results below are presented per task and separately for each age group, starting with initial performance throughout training, and followed by the pre/post-test generalisation tasks[8].

**Training**[9]

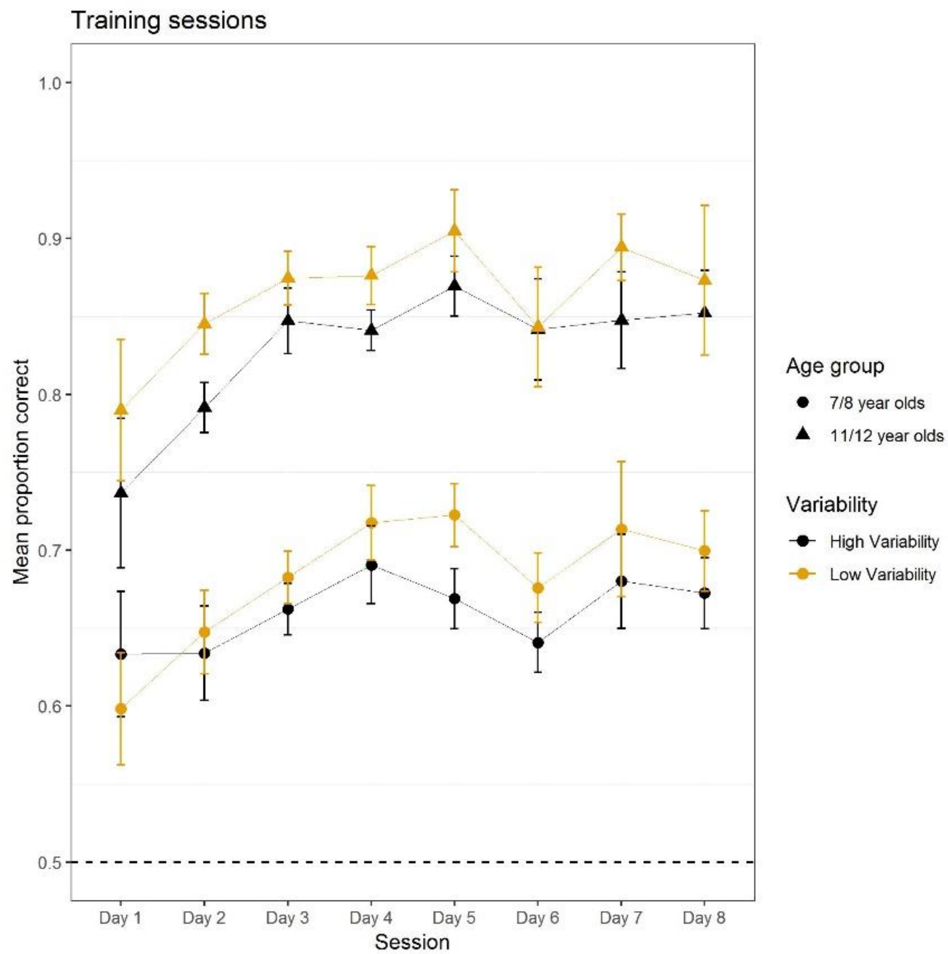


Figure 2. Response accuracy during Training. Error bars indicate 95% CI; the dashed line indicates chance.

***7-8 year-olds***

As can be seen in Figure 2, there was substantial evidence that 7-8 year-olds performed above chance in the Training task (β =0.807, SE =0.088, z =9.137, p < .001; $BF_{H(0,2.362)}$ =9.50 * $10^{16}$ [RR 0.016, >5.516]) and that they improved across sessions (β =0.033, SE =0.010, z =3.376, p =.001; $BF_{H(0,0.166)}$ =34.39 [RR 0, >5.516]). (Note that the dip in accuracy at Day 6 occurred after a weekend). However, the evidence for LV outperforming HV overall (β =0.141, SE =0.144, z =0.979, p =.327; $BF_{H(0,0.648)}$ =0.57 [RR 0, 1.116]) and for there being more improvement in LV than in HV (β =0.033, SE =0.020, z =1.640, p =.101; $BF_{H(0,0.103)}$ =1.31 [RR 0, 0.416]) was ambiguous.

***11-12 year-olds***

As seen in Figure 2, there was substantial evidence that 11-12 year-olds performed above chance in the Training task (β=2.336, SE=0.186, z=12.584, p<.001; $BF_{H(0,0.806)}$=2.03*$10^{32}$ [RR 0.116, >5.516]), and that they improved across sessions (β=0.162, SE=0.038, z=4.298378, p< .001; $BF_{H(0,0.033)}$=82.22 [RR 0, >5.516]). The evidence for LV outperforming HV overall (β=0.467, SE=0.267, z=1.748, p=.081; $BF_{H(0,1.626)}$=1.375 [RR 0, >5.516]) and there being more improvement in LV than in HV was ambiguous (β=0.029, SE=0.077, z=0.380, p=.704; $BF_{H(0,0.103)}$ =0.776 [RR 0, 0.316]).

***Age-group comparisons***

There was substantial evidence that 11-12 year-olds outperformed 7-8-year-olds overall across Training (β=1.368, SE=0.164, z=8.315, p<.001; $BF_{H(0,1.998)}$=1.34*$10^{14}$ [RR 0.1156, >5.516]) and that they showed greater improvement across training sessions (β=0.102, SE=0.023, z=4.488, p<.001; $BF_{H(0,0.083)}$=6153.59 [RR 0, >5.516]) [9].

**Discrimination**

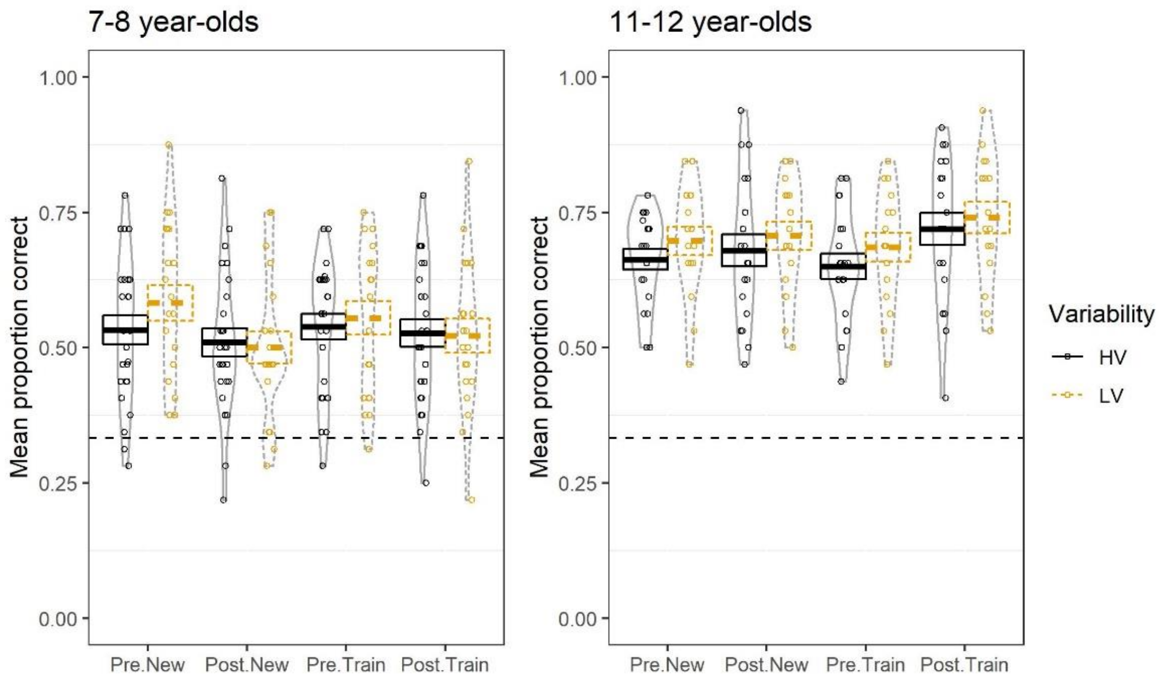


Figure 3. Response accuracy in the Discrimination task. Violins show kernel density estimation; points are individual participants. Mean performance is shown by the horizontal line with the band around it showing 95% CI; dashed line indicates chance.

***7-8 year-olds***

As seen in Figure 3, there was substantial evidence that 7-8 year-olds performed above chance in the Discrimination task ($\beta$=0.836, SE=0.097, z=8.584, p<.001; $BF_{H(0,1.687)}=1.02*10^{15}$ [RR 0, >4.830]), but substantial evidence that they did not improve overall from pre- to post-test ($\beta$=-0.158, SE=0.098, z=-1.609, p=.108; $BF_{H(0,0.205)}=0.184$ [RR 0.130, >4.8299]). Since there was no effect of session, no further interactions with session are reported.

***11-12 year-olds***

As seen in Figure 3, there was substantial evidence that 11-12 year-olds both performed above chance in the Discrimination task ($\beta$=1.687, SE=0.133, z=12.650, p<.001; $BF_{H(0,0.836)}=2.43*10^{33}$ [RR 0, >4.830]) and improved from pre- to post-test ($\beta$=0.205, SE=0.074, z=2.766, p=.006;

$BF_{H(0,1.687)}$=3.98 [RR 0.130, >2.230]). The evidence for there being more improvement in HV than in LV (β=0.050, SE=0.141, z=0.352, p=.726; $BF_{H(0,0.205)}$=0.725 [RR 0, 0.530]), more improvement in LV than in HV (β=-0.050, SE=0.141, z=-0.352, p=.726; $BF_{H(0,0.205)}$=0.445 [RR 0.030, 0.230]), and greater improvement for HV than LV for novel items only (β=0.026, SE=0.303, z=0.086, p=.932; $BF_{H(0,0.136)}$=0.936 [RR 0.030, 0.830] were all ambiguous.

**Picture Identification: Minimal Pair Items[10]**

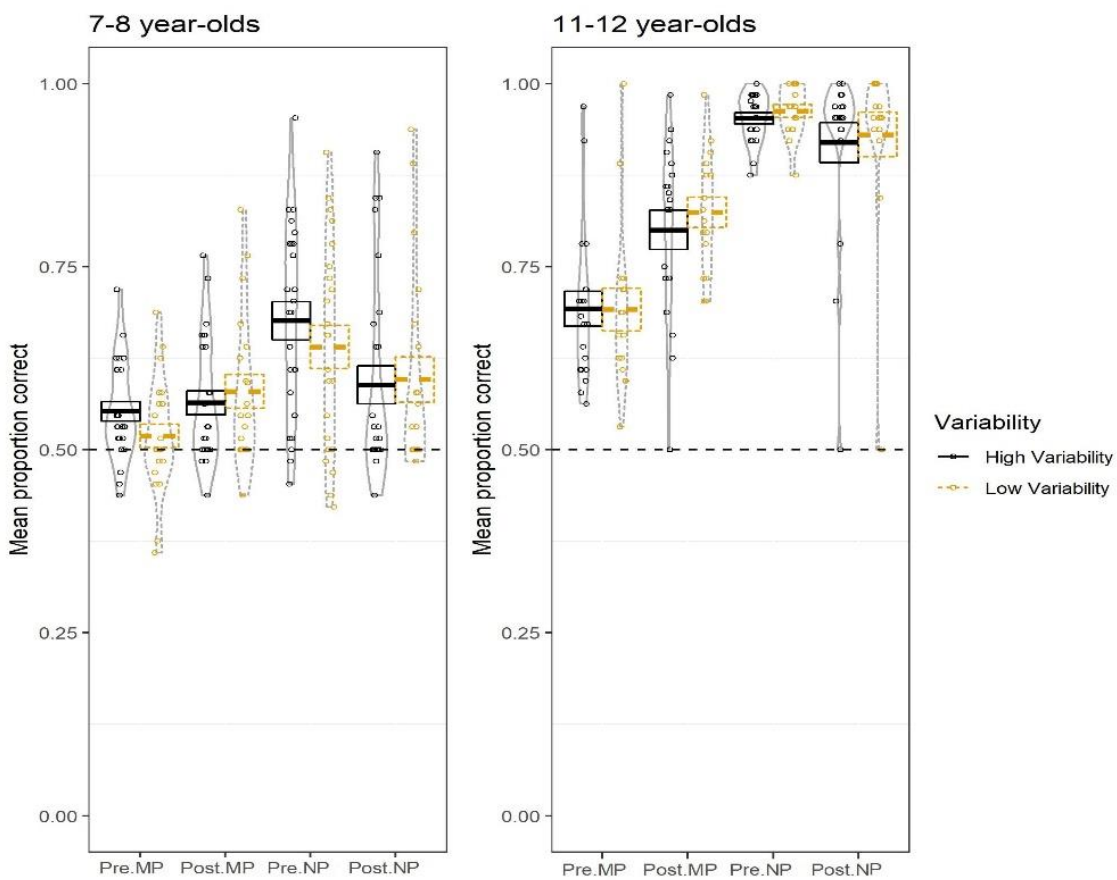


Figure 4. Response accuracy in the Picture Identification task. Violins show kernel density estimation; points are individual participants. Mean performance is shown by the horizontal line with the band around it showing 95% CI; the dashed line indicates chance. 'MP' are the minimal pair items, while 'NP' are the non-pair items.

***7-8 year-olds***

As seen in Figure 4, there was substantial evidence that 7-8 year-olds performed above chance on the minimal pair items (β=0.225, SE=0.050, z=4.504, p<.001; $BF_{H(0,1.365)}$=1835.64 [RR 0, >4.143]) but ambiguous evidence for improvement from pre- to post-test (β=0.145, SE=0.075, z=1.916, p=.055; $BF_{H(0,0.767)}$=1.174 [RR 0.263, 2.863]). Since there was not substantial evidence for an effect of session, no further interactions with session are reported.

***11-12 year-olds***

As seen in Figure 4, there was substantial evidence that 11-12 year-olds performed above chance on the non-pair items (β=1.365, SE=0.143, z=9.556, p<.001; $BF_{H(0,0.225)}=1.28*10^{14}$ [RR 0, >4.143]), and improved from pre- to post-test (β=0.767, SE=0.116, z=6.635, p<.001; $BF_{H(0,0.145)}=8.51*10^{5}$ [RR 0, >4.143]). There was substantial evidence against there being more improvement after HV than LV training (β=-0.076, SE=0.241, z=-0.316, p=.752; $BF_{H(0,0.767)}$=0.226 [RR 0.503, >4.143]), and ambiguous evidence as to there being more improvement in LV than HV training (β=-0.068, SE=0.223, z=-0.304, p=.761; $BF_{H(0,0.767)}$=0.36 [RR 0, 0.833]).

**Picture Identification: Non-Pair Items (testing vocabulary knowledge)**

***7-8 year-olds***

As seen in Figure 4 there was substantial evidence that 7-8 year-olds performed above chance on the non-pair items (β=0.5721 SE=0.078, z=7.304, p<.001; $BF_{H(0,3.300)}=1.79*10^{10}$ [RR 0, >4.143]) but substantial evidence against pre- to post-test improvement (β=-0.299, SE=0.110, z=-2.720, p=.007; $BF_{H(0,0.571)}$=0.049 [RR 0.083, >4.143]). Since there was no effect of session, no further interactions with session are reported.

***11-12 year-olds***

As seen in Figure 4, there was substantial evidence that 11-12 year-olds performed above chance on the non-pair items (β=3.300, SE=0.165, z=19.959, p<.001; $BF_{H(0,0.225)}=1.90*10^{42}$ [RR 0, >4.143]). There was substantial evidence against pre- to post-test improvement (β=-0.068, SE=0.285, z=-0.237, p=.812; $BF_{H(0,3.3)}$=0.07 [RR 0.663, >4.143]). Since there was no effect of session, no further interactions with session are reported.

**Orthography Identification**

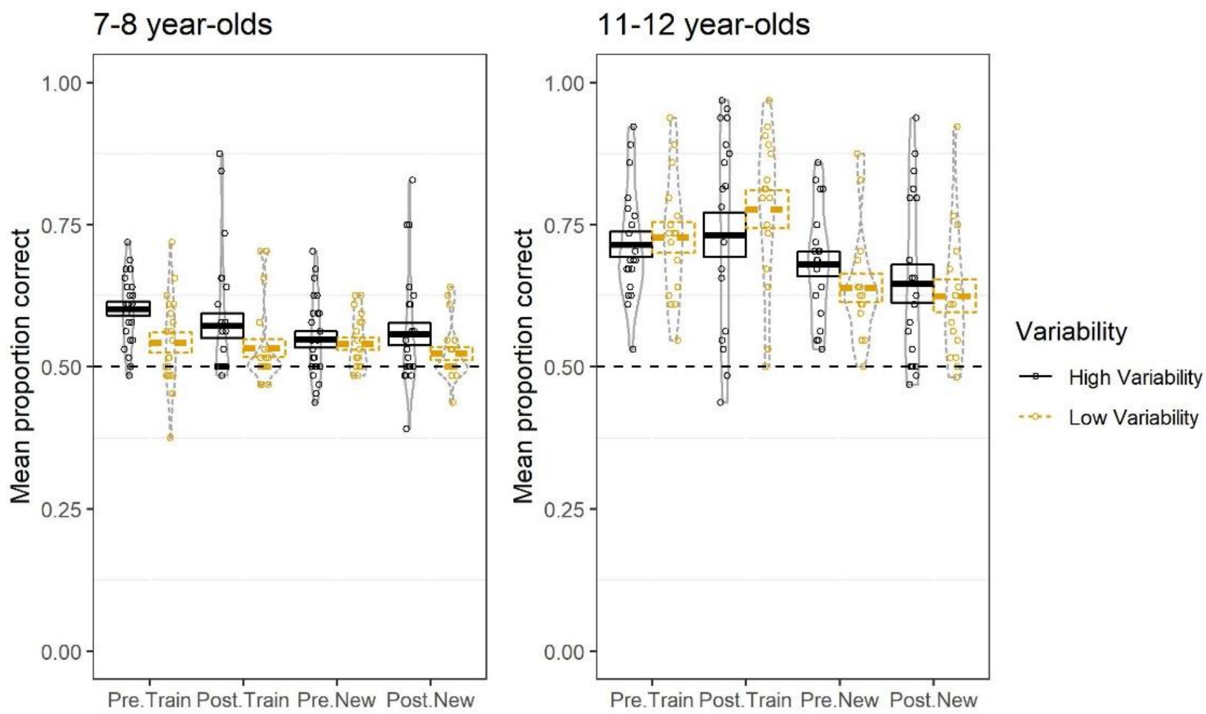


Figure 5. Response accuracy in the Orthography Identification task. Violins show kernel density estimation; points are individual participants. Mean performance is shown by the horizontal line with the band around it showing 95% CI; the dashed line indicates chance.

***7-8 year-olds***

As seen in Figure 5, there was substantial evidence that 7-8 year-olds performed above chance in the Orthography Identification task (β=0.225, SE=0.036, z=6.229, p<.001; $BF_{H(0,0.983)}$=$1.91*10^{7}$ [RR 0, >4.840]), but substantial evidence that they did *not* improve from pre- to post-test (β=-0.042, SE=0.051, z=-0.828, p=.408; $BF_{H(0,0.225)}$=0.13 [RR 0.08, >4.840]). Since there was no effect of session, no further interactions with session are reported.

***11-12 year olds***

As seen in Figure 5, there was substantial evidence that 11-12 year-olds performed above chance in Orthography Identification (β=0.983, SE=0.114, z=8.618, p<.001; $BF_{H(0,0.225)}=6.11*10^{12}$ [RR 0, >4.840]), but substantial evidence that they did *not* improve from pre- to post-test (β=0.083, SE=0.115, z=0.720, p=.471; $BF_{H(0,0.983)}=0.227$ [RR 0.67, >4.840]). Since there was no effect of session, no further interactions with session are reported.

## DISCUSSION

The present study explored whether phonetic training may be effective for primary-school-aged children. Our hypotheses concerned the role of talker variability and age in the ability to improve phonemic category identification both in training and on a range of identification and discrimination tasks conducted at pre- and post-test. Overall, this study found evidence that both age groups improved during the training task itself, although improvement was greater in 11-12 year-olds. However, although both age groups performed above chance in all of the generalisation test tasks (Discrimination, Minimal pair Picture Identification, and Orthography Identification), surprisingly 7-8 year-olds did *not* show substantial evidence of pre- to post-improvement in *any* test, with evidence for the null in every case but one. This does not mean that the 7-8 year-olds did not learn from training, as performance during training clearly increased. Instead, it means that the youngest children were unable to apply what they learnt to novel talkers in the pre/post-tests. In contrast, 11-12 year-olds improved in three out of five tasks. Since all pre/post-tests involve novel voices, this improvement indicates generalisation. However, there was no evidence that the amount of generalisation for 11-12 year-olds was greater after HV training than LV training, with one test providing evidence for the null for this hypothesis. In short, no HV benefit effects were found in any task.

In the following discussion, we first focus on why 7-8 year-olds showed no pre- to post-test improvement in any task and why there was improvement in only some tasks for 11-12 year-olds. We then consider the implications for our hypotheses about age-based differences in learning before turning to consider the findings with regard to the role of talker variability in L2 learning more generally.

**Learning in 7-8 year-olds**

The most surprising finding in this data is that although the 7-8 year-olds improve during training, indicating learning, they do *not* show improvement in any pre/post-test, and indicating no evidence of generalisation. Note that they are above chance *overall* (i.e., including at pre-test) on the Discrimination, Picture Identification and Orthography Identification tasks, indicating that, at least to some extent, they are able to distinguish the non-native vowels, and have some knowledge of the associations between English words and their meanings (the pictures), as well as some knowledge of the mappings between the vowels and their orthography. However, there is no evidence to suggest that they improve in these when tested with novel items, with evidence for the null in all cases except the minimal pair Picture Identification task – the task most like Training – where the evidence is ambiguous. This contrasts with previous studies where 7-8 year-olds showed pre- to post- improvement and generalisation after similar amounts of training (Giannakopoulou et al., 2013, 2017; Shinohara & Iverson, 2013).

One thing to consider in interpreting this lack of generalisation is whether the task demands at test were too high for 7-8 year-olds. This could be the case for Orthography Identification as 7-8 year-olds are still developing readers and writers. However, task difficulty cannot explain why we see no pre- to post-test improvement in the Picture Identification with minimal pairs and Discrimination tasks, since these have been successfully used in other research with this age-group (e.g., Giannakopoulou et al. 2017). An alternative is that the training task itself made it harder to transfer acquired knowledge to generalisation beyond the trained materials. The current task trained learners on four different vowel contrasts at the same time, whereas a majority of studies using the HVPT paradigm, and all those conducted with children, trained learners on a single phoneme contrast. Although there is evidence that training multiple contrasts is helpful for adult learners (Nishi & Kewley-Port, 2007), it is possible that this may increase complexity in a way that is detrimental for child learners. In this particular case, it seems likely that the 7-8 year-olds were unable to generalise what they had learnt during training to novel talkers and items as the knowledge that was being acquired in training itself was not yet sufficiently developed. Further consolidation through extended training sessions or training on fewer contrasts so that the overall amount of training per contrast increased might have resulted in increased generalisation for our 7-8 year-olds. Conversely, if extensive training is required on individual contrasts, this would clearly reduce the utility of the paradigm as an educational tool.

Finally, the fact that we embedded this study in a classroom-based rather than a lab-based setting, means that while we did our best to minimise noise, there were potential distractions present which might well impact younger learners. On the other hand, this setting has the benefit of higher ecological validity and the results clearly suggest limitations on the use of such paradigms in real-life educational contexts with young learners.

**Learning in 11-12 year-olds**

This age group showed pre- to post-test improvement in most tasks with the exception of the non-minimal pair Picture Identification and Orthography Identification tasks where there was evidence for the null. For the non-minimal pair Picture Identification task, this lack of improvement looks likely to be a result of ceiling effects (especially since this is primarily a measure of vocabulary learning, and improvement is found for this age group in the Vocabulary test). However, the lack of improvement for the Orthography Identification task suggests that 11-12 year-olds may not actually use the orthography presented during training. If so, this contrasts with existing findings showing that for Dutch adult learners of English, spelling is especially helpful initially when learning words that rely on non-native phoneme contrasts (Escudero et al., 2008).

Theoretically, it seems plausible that mapping to orthographic representations could help with developing phoneme categories, since they provide an invariant representation of those categories across instances. However, it is an open question whether this is actually helpful for children with less reading experience. Indeed, Shinohara & Iverson (2021) also found a lack of improvement in their youngest age group (6-8 year-olds) and attributed this to difficulty in using orthography in training. That said, it is possible that orthography *was* helpful in our study, even though children did not show learning when tested on orthography in isolation: it might have an additive benefit when presented in combination with the pictures. To shed light on this, future research could include a pre/post-test which *combines* orthography and pictures to check for better performance in comparison to picture/orthography-only tasks. On the other hand, there is evidence from some studies that the inclusion of orthography can actually *harm* learning. Showalter & Hayes-Harb (2015) showed negative effects of adding orthography for word learning when the orthography was completely unfamiliar, and Hayes-Harb et al., (2018) found a negative influence of orthography on L2 speech production (see Hayes-Harb & Barrios (2021) for a more extensive review). Whether adding orthography actively *harms* learning in this paradigm could be tested by comparing the learning in the current study with a condition with pictures-only training.

**Age-based differences in learning**

It is clear that there was no age effect in favour of younger learners: younger learners did not benefit more from training and did not show more improvement during training. In contrast, we find clear evidence that not only do older children perform better at the start of training (likely as a result of their greater experience with English), but that they also learn more during training. This contrasts with previous findings that have shown younger is better (e.g., Giannakopoulou et al., 2013). Shinohara (2014) also found an age benefit for 8-12 year-olds compared to adults, however 6-8 year-olds did not show any pre- to post-test improvement, a pattern similar to the current study.

Older learners outperforming younger learners is not unprecedented in language learning research: Huyck & Wright (2010) found that adolescents showed much less, or even no, improvement in auditory perceptual learning tasks compared to adults who did improve, despite all adolescents starting out within the adult range of scores at the start of training. Moreover, Snow & Hoefnagel-Höhle (1978) found that younger learners consistently showed less improvement in various areas of natural L2 learning compared to older children and adults. This suggests that there may be more to this apparent benefit of younger age, and that it may depend on more than just age factors (cf. Flege & Bohn, 2021).

It is possible that any inherent benefits for younger learners may have been offset by task difficulties. We also cannot rule out that our 7-8 year-olds might catch up and eventually overtake 11-12 year-olds given more training (note that even after 8 training sessions, their performance did not reach the starting point of the 11-12 year-olds in this study). On the other hand, it is notable that although starting younger appears to be beneficial for long-term phonetic learning in naturalistic environments (Flege, 1995), there is actually relatively little direct experimental evidence in the literature that younger is better for phonetic training. One possibility is that older children benefit more from training because they have a better starting point: if phonetic training leads to more accurate mappings from an existing L1 to an L2 category (cf. Iverson & Evans, 2009), then young children's L1 categories might not be well-developed enough to support learning in this way. As children grow older and their L1 categories become more stable, their learning strategies might change to more closely resemble those of adults such that they show a benefit from phonetic training.

Another possibility relates to differences in the amount of exposure to English between the two age groups. If the younger children start out having less exposure than the older children, then this may influence how well they do in certain tasks. While older children might be fine-tuning their knowledge of the words, younger children with no prior knowledge to access, might be unable to do so. Note that prior knowledge of the words in and of itself would not necessarily be helpful. Given that the children are in a Dutch context, part of their exposure to English will have been in a Dutch-accented context where the phoneme contrasts of interest might not have been correctly or consistently distinguished. This could potentially be detrimental if the mental representation of the words were inaccurate for the vowel contrasts in question. Theoretically, this could put older learners at a disadvantage as they would need to adjust fixed representations. In contrast, younger learners, without pre-existing lexical representations, could start “from scratch”. One possibility in the current study is that the relatively small amount of learning and familiarity in the 11-12 year-olds is enough to make the learning tasks easier but not enough to lead to inaccurate representations which are too entrenched and therefore resistant to change.

On the other hand, there was a relatively small difference in the amount of English learning by our participants – although 7-8 year-olds had no formal training, 11-12 year-olds had had one session a week in basic English for a year. However, our reading of the literature is that there is no clear pattern as to whether learners with more L2 experience show greater or lesser improvement. For example, Giannakopoulou et al. (2017) found that adults with on average over 10 years of exposure showed less improvement than children with 1-2 years English education, but Shinohara & Iverson (2021) found that adults with many years of exposure through schooling improved more than 8-10 year-olds with no formal English training, but *less* than 16-18 year-olds with four to six years of formal English training. It is thus difficult to draw strong conclusions about the role of previous L2 exposure, or to pull this apart from age at the time of testing.

When considering the impact of possible age-based differences in the phonetic training paradigm, it is worth considering that others might have found no younger-is-better effect, but that due to the file drawer problem these null findings have not been published (Franco et al., 2014). In addition, traditional statistical methods make it difficult to interpret null effects where these are reported (e.g., Heeren & Schouten (2010) who found no effect of age on learning). The current work makes a contribution in using Bayes factors to clearly show evidence *against* a “younger is better” advantage for learning in training. While the study aimed to investigate similar effects in the pre/post-test tasks, the lack of improvement across these tasks for the younger age group means this comparison would not have been statistically informative beyond confirming that one group did indeed not learn at all. Future studies could compare different age-groups using a variety of training and test tasks, to explore if and when an age effect is apparent, and whether or not this is affected by the types of task and materials. Nevertheless, from an educational perspective, the current results speak against the utlitity of this type computerised training in children as young as 7, at least in its current form, and more generally caution against the mantra of “younger is better” when it comes to foreign language teaching.

**Variability**

There are theoretical reasons to expect a variability benefit in the current paradigm, since encountering variability should help listeners to learn to implicitly ignore irrelevant speaker-specific cues, tune into cues relevant for the phonetic contrasts, and generalise learning to new stimuli. However, an HV benefit was *not* found for 11-12 year-olds (the only age-group who showed any evidence of generalisation and where we can thus draw conclusions about variability) in the current study, with evidence for the null in one of the tests (Minimal Pair Picture Identification). Note that due to a lack of pre-to-post improvement we did not test for a benefit specifically for test stimuli involving novel items (cf. Evans & Martín-Alvarez, 2016). One possible explanation for the lack of variability benefit is that the task demands were too high even for our 11-12 year-olds. One possibility is that the added complexity of dealing with multiple talkers in training, in addition to the complexity of training multiple vowels simultaneously as discussed above, combined to make the HV training task too cognitively demanding for the children and that this offset any potential benefit of talker variability. This explanation is similar to that in Sinkeviciute et al. (2019), where younger children showed an LV benefit in (vocabulary) training but no HV advantage for generalisation, in contrast to adults who showed an HV advantage for generalisation. However, we note that in the current study we did *not* find evidence of an LV benefit in training for either age-group (the evidence was ambiguous). We also found no LV benefit in generalisation – here with evidence for the null in one of the tasks, failing to replicate Giannakopoulou et al. (2017), who found this in 7 year-olds in both training and generalisation.

It is also conceivable that other factors might lead to the lack of variability effect. For example, we could speculate that distractions in our classroom-based rather than lab-based context could potentially impact on the difficulty of benefiting from variable materials. When interpreting the findings, it is also important to recall that HV benefits for generalisation are *not* consistently reported in phonetic training studies even in adults (as noted in the introduction and discussed more extensively in Brekelmans et al., 2022). Future studies must unravel the extent to which variability effects in phonetic learning and generalisation – and the lack thereof – are impacted by age and other factors.

## CONCLUSION

The current study trained native Dutch 7-8 year-olds and 11-12 year-olds on four English vowel contrasts using a two-week computerized phonetic training paradigm embedded in a classroom context. While both age groups showed improvement during training, 11-12 year-olds showed greater improvement and, critically, there was evidence that 7-8 year-olds were *not* able to generalise this acquired knowledge to any of the post-test tasks using novel speaker. This suggests age-based differences, and limitations in the application of phonetic training with 7-8 year-olds in this context. Crucially, where generalisation did occur, this study did not find that it was increased by talker variability in the input.

**NOTES**

[1] A pilot with these materials with 27 16 year-olds found that they were at ceiling even at pre-test.

[2] Note that the pre-registration also includes a production task, phonological working memory test, and category boundary task. We had planned to compare production and comprehension (following the different findings in Evans & Martín-Alvarez, 2016) but were not able to do this due to practicalities when testing in school. Due to data loss, these tasks are not reported as part of the current study. A Vocabulary task was also pre-registered; for space reasons full methods and results for this task are reported on OSF.

[3] Note that there are places where the literature might lead us to expect no effect (e.g. no variability effect in word learning in 7-year-olds). However, Bayes Factors require us to lay out a positive hypothesis, which we then test against the null.

[4] The intercept *i* in these models represents the grand-mean (in log odds) so to get the difference *d* from chance *c,* we can compute *d=i-c*. Where chance is 50%, log odds is 0 so we have *d=i,* where chance is 1/3, this is -0.69 in log odds, so our *d=i - -0.69.* This value *d* is the mean difference that we feed into our Bayes Factor computation (along with the standard error *se* associated with that intercept); we get p-values from the z-score *z=d/se*.

[5] One reviewer suggested instead of doing the analyses per age group separately first and then doing a direct comparison of age groups, we do one large analysis with nested fixed effects (as explained in Schad et al., 2020). We tried this and it gave the same results, but some of the models had convergence issues. Given this and our desire to stay as close to the pre-registration as possible, the results reported in the paper are those proposed in the pre-registration.

[6] As preregistered, for each set of analyses, before any models were run, performance was plotted by vowel contrast to ensure performance on the control contrast FLEECE-THOUGHT was not near ceiling (in which case the data would have been removed); this was never the case. No

further analyses were done for this factor though data are available on OSF. Full analyses run without the FLEECE-THOUGHT control contrast in the data can be found on OSF: the overall pattern of results remained the same.

[7] Note that the mean encodes what we expect to be the most probable effect size. While at first glance it might seem this should be set to the predicted effect size, this would make smaller effects symmetrically less probable, whereas in psychology our effect sizes are generally small and smaller effects are more probable than larger ones.

[8] Note our analyses deviate from the preregistration due to unexpected null effects in our results. While the proposed analyses planned to test interactions with *session* and *variability* in all pre/post-tests, we did not anticipate null effects of learning. If there is no effect of learning, it makes no sense to see if the effect of learning differs depending on the variability condition, as no evidence for an effect was found in the first place. Since these analyses were pre-registered, we did conduct these and they are reported on OSF (https://doi.org/10.17605/OSF.IO/BGDXP) for completeness. In almost every case, they yielded ambiguous evidence, and where they did not, they yielded substantial evidence for the null. In the current paper, if no substantial evidence was found for an effect of session (i.e. no improvement from pre- to post-test was seen), no further follow-up interactions were reported. The preregistration further proposed age comparisons between the 7-8 year-olds and 11-12 year-olds on every pre/post-test, to test for differences in the amount of generalisation shown in the age groups. Given that the 7-8 year-olds consistently show a lack of improvement on pre/post-tests, we do not think it sensible to report age comparisons looking at the amount of improvement in generalisation if one of the two groups showed none to start with. For this reason, age comparisons are only reported for training, where learning across sessions was found in both age groups. Since these analyses were pre-registered, they were conducted and are reported on OSF for completeness. In almost every case, there was evidence for older learners outperforming younger learnings, and often also for older learners improving more than younger learners. Where this was not the case there

was often evidence *against* younger learners outperforming older learners but ambiguous evidence in the opposite direction.

[9] A reviewer pointed out that change over sessions in the training model might be better modelled as a curve rather than a linear predictor. Although this was not pre-registered, we tried running identical models to those presented here with a quadratic instead of a linear slope. The only model that would converge was an intercept-only model without vowel contrast covariates. This showed the same pattern of results as the model reported in this paper.

[10] We also tested the reverse hypothesis that younger learners improved more than older learners and there was substantial evidence for the null ($\beta$=-0.102, SE=0.023, z=-4.488, p<.001; $BF_{H(0,0.083)}$=0.049 [RR 0, >5.516]).

[11] The pre-registration proposed analysing the data in the same way as other pre/post-tests, with item novelty as a factor. However, in hindsight, the contrast between the minimal pair trials and the non-minimal pair trials is quite different from the item novelty manipulation, as all words are familiar in the non-pair condition. This difference is reflected in rather different performances on the minimal pair items compared to the non-pair items. Therefore, we felt it more appropriate to analyse the minimal pair items separately from the non-pair items for each age group.

## REFERENCES


Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child development, 77*(6), 1698-1716. https://doi.org/10.1111/j.1467-8624.2006.00968.x

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

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

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects39odellingg 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

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

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

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

Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer* (Version 5.4.17). http://www.praat.org/

Bolker, B. (2019). *GLMM FAQ*. http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#should-i-treat-factor-xxx-as-fixed-or-random

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. I. (1999). Training Japanese listeners to identify English/r/and/l: Long-term retention of learning in perception and production. Perception & psychophysics, 61(5), 977-985. https://doi.org/10.3758/BF03206911

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

Braginsky, M. (2018). *ggpirate: pirate plots for ggplot2*. https://github.com/mikabr/ggpirate

Brekelmans, G., Lavan., N., Saito, H., Clayards, M. & Wonnacott, E. (2022). Does high variability training improve the learning of non-native phoneme contrasts over low variability training? A replication. *Journal of Memory and Language, 126.* 104352. https://doi.org/10.31234/osf.io/gs7y9

Cebrian, J., & Carlet, A. (2014). Second-Language Learners' Identification of Target-Language Phonemes: A Short-Term Phonetic Training Study. *Canadian Modern Language Review*, *70*(4), 474–499. https://doi.org/10.3138/cmlr.2318

Collins, B., & Mees, I. M. (2003). *The Phonetics of English and Dutch* (Fifth revised edition). Leiden: Brill.

Crespo, K., & Kaushanskaya, M. (2021). Is 10 Better than 1? The Effect of Speaker Variability on Children's Cross-situational Word Learning. *Language Learning and Development*, *17*(4), 397–410. https://doi.org/10.1080/15475441.2021.1906680

Crespo, K., Vlach, H., & Kaushanskaya, M. (2023). The effects of bilingualism on children's cross-situational word learning under different variability conditions. *Journal of Experimental Child Psychology*, *229*, 105621. https://doi.org/10.1016/j.jecp.2022.105621

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

Dienes, Z. (2008). Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference. Palgrave Macmillan.

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*(July), 1–17. 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.), *Behavioral Methods in Consciousness Research* (pp. 199–220). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199688890.003.0012

Dienes, Z. (2017). *Bayes Factors: Context, Principles, Criticism.* Retrieved from https://www.youtube.com/watch?v=g9bIfZ4KqCQ

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

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

Escudero, P., Hayes-Harb, R., & Mitterer, H. (2008). Novel second-language words and asymmetric lexical access. *Journal of Phonetics*, *36*(2), 345–360. https://doi.org/10.1016/j.wocn.2007.11.002

Eurostat. (2019). *Foreign language learning statistics*. Retrieved August 15, 2019, from Eurostat: Statistics Explained website: https://ec.europa.eu/eurostat/statistics-explained/index.php/Foreign_language_skills_statistics

Evans, B. G., & Martín-Alvarez, L. (2016). Age-related differences in second-language learning?A comparison of high and low variability perceptual training for the acquisition of

English /i/-/ɪ/ by Spanish adults and children. *Proceedings of New Sounds.: 8th International Conference on Second Language Speech.* Aarhus, Denmark. Retrieved from https://conferences.au.dk/newsounds2016/

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., & Bohn, O. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress* (pp. 3-83). Cambridge: Cambridge University Press. https://doi.org/0.1017/9781108886901.002

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

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. Science, 345(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Fuhrmeister, P., & Myers, E. B. (2020). Desirable and undesirable difficulties: Influences of variability, training schedule, and aptitude on nonnative phonetic learning. Attention, Perception, & Psychophysics, 1-17. https://doi.org/10.3758/s13414-019-01925-y

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

Giannakopoulou, A., Uther, M., & Ylinen, S. (2010). Phonetic cue-weighting in the acquisition of a second language (L2): evidence from Greek speakers of English. *Achievements and Perspectives in SLA of Speech: New Sounds 2010, 1*(1), 91–201.

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*(2), 201–218. https://doi.org/10.1177/0265659012467473

Hayes-Harb, R., Barrios, S. (2021) The influence of orthography in second language phonological acquisition. *Language Teaching*, *54*(3). 297-326. https://doi.org/10.1017/S0261444820000658

Hayes-Harb, R., Brown, K., & Smith, B. L. (2018). Orthographic Input and the Acquisition of German Final Devoicing by Native Speakers of English. *Language and Speech*, *61*(4), 547–564. https://doi.org/10.1177/0023830917710048

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

Huckvale, M. (2016). *ProRec: Prompt & Record* (2.3). http://www.phon.ucl.ac.uk/resource/prorec/

Huyck, J. J. & B. A. Wright. (2010). Late maturation of auditory perceptual learning. *Developmental Science, 14*(3), 614-621. https://doi.org/10.1111/j.1467-7687.2010.01009.x

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

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

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), 148-203. http://dx.doi.org/10.1037/a0038695

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

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

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. Journal of experimental psychology. Learning, memory, and cognition, 15(4), 676–684. https://doi.org/10.1037//0278-7393.15.4.676

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

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

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

Nishi, K., & Kewley-Port, D. (2008). Nonnative Speech Perception Training Using Vowel Subsets: Effects of Vowels in Sets and Order of Training. *Journal of Speech, Language, and Hearing Research, 51*(6), 1480–1493. https://doi.org/10.1044/1092-4388(2008/07-0109)

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

Powell, M. J. D. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. *NA Report NA2009/06*, 39. https://doi.org/10.1.1.443.7693

R Core Team. (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.r-project.org

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

Sadakata, M., & McQueen, J. M. (2014). Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training. Frontiers in psychology, 5, 1318. https://doi.org/10.3389/fpsyg.2014.01318

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, 110, 104038. https://doi.org/10.1016/j.jml.2019.104038

Shinohara, Y., & Iverson, P. (2021). The effect of age on English /r/-/l/ perceptual training outcomes for Japanese speakers. *Journal of Phonetics*, *89*, 101108. https://doi.org/10.1016/j.wocn.2021.101108

Showalter, C. E., & Hayes-Harb, R. (2015). Native English speakers learning Arabic: The influence of novel orthographic information on second language phonological acquisition. *Applied Psycholinguistics*, *36*(1), 23–42. https://doi.org/10.1017/S0142716414000411

Silvey, C., Dienes, Z., & Wonnacott, E. (2024). Bayes factors for logistic (mixed-effects) models. *PsyArXiv.* https://doi.org/10.31234/osf.io/m4hju

Sinkevičiūtė, R., Brown, H., Brekelmans, G., & Wonnacott, E. (2019). The role of input variability and learner age in second language vocabulary learning. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263119000263

Snow, C. E. & M. Hoefnagel-Höhle. (1978). The Critical Period for Language Acquisition: Evidence from Second Language Learning. *Child Development, 49*(4), 1114-1128. https://doi.org/10.2307/1128751

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

Wang, Y., & Kuhl, P. K. (2003). Evaluating the "critical period" hypothesis: perceptual learning of Mandarin tones in American adults and American children at 6, 10 and 14 years of age,". In Proceedings of the 15th International Congress of Phonetic Sciences (pp. 1537-1540). Barcelona, Spain: Futurgraphic. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_1537.html

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag.

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

Wong, J. (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 *Proceedings of the 14th Australasian International Conference on Speech Science and Techniology. Sydney, Australia* (pp. 3-6). https://assta.org/proceedings/sst/SST-12/SST2012/PDF/AUTHOR/ST120021.PDF

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