Phonetic vowel training for child second language learners: the role of input variability and training task

Gwen Brekelmans

A thesis submitted for the degree of

Doctor of Philosophy

Department of Language & Cognition

Division of Psychology and Language Sciences

University College London

2020

Declaration

I, Gwen Brekelmans 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 and foremost, thank you so much to Liz Wonnacott for being an amazing supervisor. Thank you for your help with pretty much anything, for all your hard work, and your statistical knowledge; I've learnt so much over the past four years. Thanks also for littering my writing drafts with helpful comments and restructuring suggestions, which have made this thesis a lot easier to follow. Any remaining structural randomness is entirely my own, and any typos have earned their right to be printed by managing to slip under the radar. Many thanks as well to Bronwen Evans for being a fantastic second supervisor: thank you for all your phonetic wisdom, helpful advice, inspiration, and incredible kindness.

I'm eternally grateful to all the participants, adults and children alike, with extra thanks to the clever and wonderful children for all the hair I braided, laces I tied, stickers I got to hand out, and stories I was told, making testing anything but boring. Particular thanks also to all the teachers and head teachers at the schools I tested at, who let me wreak havoc on their schedule by taking children out of class in between sports days, Roman History projects, and play performances of Matilda. Thanks also to Anna, Cathy, Claire, Lewis, Matthew, Natasha, Rachel, Harrie, Karin, Lex, Maria, and Riek for being the voices in my experiments, and in doing so eliciting delightful comments from some very puzzled 7-year-olds.

I'm very grateful to Stuart Rosen and Kathleen McCarthy for providing me with the category boundary task for Study 2, even if I didn't end up using it because of spectacularly timed technical difficulties during the first day of testing. Thanks to Jason Shaw for providing the code for the production analysis of Study 2, and to Andrew Clark for the LaTeX code to make the vowel diagrams.

Many thanks to my fellow language learning lab members: Hanyu Dong, Anna Samara, Cat Silvey, Daniela Singh, Maša Vujović, and Liz Wonnacott, for being such a supportive lab. Thanks also to the various other labs and journal clubs I was adopted into, and to the wonderful people in Chandler House who made creating this thesis a little easier, in particular to Anna Casey, Andrew Clark, Richard

Jardine, Nadine Lavan, Merle Mahon, Carolyn McGettigan, Caroline Newton, David Newton, Steve Newton, Courtenay Norbury, Rachel Rees, Tim Schoof, Gisela Tomé Lourido, Outi Tuomainen, Rosemary Varley, and Vitor Zimmerer.

Thanks to all the wonderful people in the LangCog PhD office for making it such a fun place to work, particularly to Ria Bernard, Lena Blott, Claudia Bruns, Sabrina Mahmood, Vanessa Meitanis, Claire Murphy, Jo Saul, Daniela Singh, Anna Volkmer, Maša Vujović, and Lydia Yeomans. Special thanks to Vanessa, Anna V, Lena, and Claudia for many more adventures outside of the office as well.

A million thanks to Marina for your support and shared 'phon phon' enthusiasm. Thanks so much to Nadine for being equally baffled by the variability literature, and happy to rant about it (and many other things since) over tea. Thanks also to Giulia, Max, Vanessa, and Anna E for the regular lunch breaks, Shiran and Anna K for the coffee chats, and Julie for the 'last PhD year panic coffees' that never happened but the many dinners that did.

Many thanks to Andrew, Anna V, Carolyn, Cat, Isabel, Nadine, and Vanessa for regularly letting me shove books at you and indulging my passionate raving about whatever latest quirky scifi/YA novel had struck my fancy, and to all the writers of those same books for providing a very welcome distraction to my research.

Huge thanks to Željka and Tori for being the best RavenPuffIn housemates, to Alessa, Bas, Cüyén, Eliza, Maya, and Mengfan for the fun and friendship despite the multiple countries between us, and to Isabel for the theatre breaks. *Kiitos* especially to Sanni for keeping me sane and being there to answer the most random questions at all times, from ponderings on the pronunciation of 'ooh' versus 'uhh', to grammar checking odd sentences (including these acknowledgements).

And finally, to my parents Riek and Harrie, my grandma Miet, and my brother Lex, for their loving support throughout the whole PhD: *hartstikke bedankt, ik had het niet zonder jullie gekund! De laatste twee jaar zijn zeker niet makkelijk geweest maar het is ons toch maar gelukt. Pap, ik weet dat je keitrots zou zijn geweest.*

Vur ons pap

2 October 1959 – 24 December 2017

Abstract

Acquiring a second language speech contrast that does not exist in the native language is often difficult. High variability phonetic training (HVPT) is a well-established method used to train learners on specific non-native phoneme contrasts: it critically uses high variability (HV) input after earlier attempts using low variability (LV) input had proved unsuccessful. HVPT has since been successfully applied in many different adult studies. However, there is no consensus on the effect of input variation on children's learning of non-native phoneme contrasts. This thesis aims to further investigate the effect of input variability on phonetic training for children, and examining whether they show the same HV benefit which has been argued to hold for adults.

In the first set of studies, native English speaking adults and children were taught Dutch vowels in a single computerised training session, during which they received either HV or LV input. Additionally, the traditional HVPT paradigm was adapted to see if mapping vowels to orthography-like symbols representing phoneme categories was more or less effective than a vocabulary training method without such representations. Learning was stronger with training most akin to vocabulary learning, particularly for children, suggesting a benefit for a more meaningful learning context. Crucially, there was no evidence of a HV benefit for either children or adults.

The second study was a two-week training study in which Dutch children of two age groups were trained on Standard Southern British English vowel contrasts. Since picture-based training had proved beneficial, this study combined both orthography and pictures in training. Potential effects of HV or LV input in training were investigated using a pre/post-test design. Older children outperformed younger children throughout, and again no evidence for a variability benefit was found. This indicates children might not benefit from high input variability.

Impact Statement

The majority of the world's population learns at least one language in addition to their native language(s). However, learning to perceive and produce the sounds of a non-native language is often difficult. Part of this is due to the complexity of speech: the input a learner receives is inherently variable, and learning which cues to pay attention to takes time and practice. The studies in this thesis investigate how children acquire second language speech in the context of a phonetic training paradigm. The research in this thesis particularly focusses on the role of variability and the type of training task used in acquiring non-native speech sounds. Although there is a sizable literature with adult learners in this area, few previous studies have investigated these questions working with children. The research in this thesis contributes towards a better understanding of the role of variability in second language speech learning, as well as providing a foundation for future research into the role of variability in second language speech learning in children.

This thesis also makes an important methodological contribution by being the first in the phonetic training literature to use Bayes Factors as the key method of inference throughout the thesis. This is an advance because with traditional frequentist methods of inference, it is not possible to say whether a non-significant result actually provides evidence *against* the set hypothesis, or is ambiguous. Bayes Factors differentiate these two possibilities. Further, the second study presented here was pre-registered, which means the analyses and hypotheses were planned

in advance and the reader can compare the planned analyses and hypotheses with the ones presented in this thesis. This is important in preventing publication bias which is known to be a problem in the field.

This thesis also has more practical implications for second language learning in naturalistic settings. The finding that variability is not beneficial for children learning a second language means that schools might want to focus on providing varied content while potentially not needing to put as much effort and resources in providing content from a large number of talkers. Rather, this thesis suggests that the choice of task used for learning seems to play a more important role. This also extends to language learning software development and computer assisted language learning, where the choice of task used in the programs could prove more influential more than the amount of talker variability provided in the input. However, for all of these implications, it is important to keep in mind that this thesis is limited to Dutch learners of English and English learners of Dutch, learning specific vowel contrasts, and that caution should be exercised in extending them to other groups of learners or other types of training.

Table of contents

List of tables

List of figures

List of vowel diagrams

Chapter 1. Introduction

The majority of the world's population speaks more than one language, and can be said to be bi- or multilingual. Learning a second or foreign language is thus a very common yet important skill. To illustrate, nearly two thirds of working-age adults in the EU reported knowing at least one foreign language (Eurostat, 2019b). Second language (L2) learning is not just restricted to adults: children make up a large portion of L2 learners as well (on average 80% of children in EU primary education learn a foreign language, and 59% of children in upper secondary education in the EU learn two or more foreign languages (Eurostat, 2019a)). This prevalence of learning second or foreign languages has sparked a prolific area of research, though notably, child L2 learning is a rather neglected area of study compared with adult L2 learning. This research aims to investigate a key area of difficulty which people experience when acquiring a second language: mastering the non-native speech sounds. This particular issue will be the focus of this thesis.

Acquiring the sounds of a non-native language after first having acquired another language (or languages) is often difficult. However, there is by now a well-established consensus in L2 speech research that second language learners do not have problems acquiring *all* sounds of an L2, but have particular difficulty mapping those sounds that do not occur in their native language (L1) onto their existing phoneme representations (see Strange (1995) for a review). A large part of this can be explained through transfer from the L1 that occurs when attempting to acquire

the speech of an L2. One reason that this transfer might occur is because speech is inherently variable, so speakers of a particular language need to learn which of the many features and cues that vary in their input are linguistically relevant, and which are not. However, a learner will observe the L2 through the knowledge of their L1, thereby biasing and filtering the input they receive. This means that L2 learners might not start out paying attention to the right cues if their L1 requires them to focus on a different set of cues. Learning to retune perception to be able to successfully acquire the L2 speech sounds takes time, practice, and even then might never be entirely native-like.

Perhaps the most influential theory in L2 acquisition is the critical period hypothesis (Lenneberg, 1967), which states that there is an ideal time window, usually up until around puberty, within which people can best learn language. Although the original theory concerned L1 acquisition, the idea of a critical period has also been extended to L2 acquisition, where it implies that fully acquiring a second language after puberty would be hard. Looking at L2 acquisition in a naturalistic setting, Johnson & Newport (1989) showed that native speakers of Korean or Chinese who arrived in the US at an earlier age showed better performance on a English grammaticality judgement task. This study also appeared to find evidence for a sharp decline in ability at puberty (in line with the original "critical" period hypothesis): there was evidence for a gradual decline in performance as the age of learning increased up to puberty, but there was no systematic relationship between age of acquisition and L2 performance if learning

started after puberty. This study has since been replicated and extended in numerous languages with different materials (see e.g. Bialystok & Miller (1999); Birdsong (1999); Birdsong & Molis (2001); Boxtel, Bongaerts, & Coppen (2005); Dąbrowska (2019); DeKeyser (2000); DeKeyser, Alfi-Shabtay, & Ravid (2010); Vanhove (2013)), and recently with a very large sample of nearly 670,000 participants by Hartshorne, Tenenbaum, & Pinker (2018). In general, all of these studies find evidence for a negative correlation between age of acquisition and ultimate ability, although there is some debate whether there is a period of sharp decline, or whether more gradual loss occurs throughout the lifespan. However, these studies all investigate the critical period in the context of grammar and syntax abilities, while the focus of this thesis is on L2 speech perception.

When looking at L2 speech research, there is also a negative correlation between the age at which a learner started learning, and their L2 phonological attainment. Importantly, here there is clear evidence of gradual decline over the lifespan rather than a sudden cut-off point as proposed in the original hypothesis. For example, Flege (1995) summarises results from various studies with adult L2 learners of various native languages, showing learners who started acquiring the L2 generally performed better in both L2 perception and production, and Ioup (2008) reviews studies focussing on comparing the production of child L2 learners to those who started learning as adults through means of both native speaker judgements and acoustic measures. The general take-home message from both of these overviews is that learners who started acquiring an L2 at a younger age tend to outperform

those who started later in terms of the phonological acquisition of the L2. The gradual nature of this decline suggests that acquiring the sounds of an L2 is possible across the lifespan, but that it might become harder over time, and that children might benefit from starting early. This is generally assumed to be because child learners are less set in their L1 as they might not have completely established its phonetic categories (Hazan & Barrett, 2000), although it is difficult to distinguish this from a more general decline in neural plasticity.

If children's L1 categories are indeed less established, this would suggest that they should be more flexible in their ability to handle the rich variability in speech, both in the form of within-category variation and between-talker variation. Though such variability is inherently difficult, there is evidence in adult literature that speaker variability is needed in training materials meant to improve speech perception through means of phonetic training, but less is known about children. One thing to note is that not all variability that occurs in speech will be useful to a listener. Generally speaking, variability can be divided into structured variation, when the variation has some sort of underlying pattern or structure to it, and truly random variation that is essentially noise in the signal. An example of structured variation would be variation in f0, which is linked to a speaker's vocal tract length and thus indirectly indicates the speaker's body size. It is the information encoded in this type of structured variation that listeners could use to make inferences about their speaker and could use in speech perception. It is for this reason that variability in the number of speakers in particular can be of use to listeners, as this type of

variability provides them with more evidence as to what part of the variation is truly random and what part of the variation is actually structured. The key focus of this thesis is to investigate phonetic training in children, asking the question: is training with high variability materials, specifically materials produced by multiple talkers rather than a single talker, as useful for child learners as it appears to be for adults? In addition, this thesis also aims to shed light on other aspects of the methods used in phonetic training which can benefit children's learning.

The aim of this chapter is to provide an overview of current research in the role of variability in second language speech learning. The first section will begin by reviewing existing literature concerning the role of variability in first language speech perception, considering different types of variability and what kind of effect they have on speech perception, before turning to second language speech perception in the second section. The third section moves to discuss the literature on high variability phonetic training, which is the key focus of the current thesis. The fourth section reviews a related literature on second language vocabulary learning, which is also related to the current thesis since some of the studies involve tests of vocabulary. Following this, the fifth section is a brief discussion of some of the methodological concerns in the field of phonetic training and in psychology more broadly, which are relevant to the current thesis. The final section provides a roadmap for the rest of this thesis.

1.1 Variability in first language speech perception

It has long been questioned how listeners deal with the wealth of variation that occurs in natural speech. Critically, instances of the same phoneme vary acoustically but perceptually still fall within the same phoneme category. For instance, English /p/ and /b/ differ in Voice Onset Time (VOT), where longer VOT leads to the perceptually voiceless sound /p/, while the same sound with a shorter VOT is perceived as the voiced /b/, but individual instances of these phonemes fall along a continuum of VOT length. Fine-grained differences in VOT have been shown to affect lexical access, leading to more lexical competition the closer the VOT comes to the category boundary (McMurray, Tanenhaus, & Aslin, 2002). This effect has been found to extend to various different tasks with different task demands, such as lexical decision and phoneme decision with both real words and non-word syllables (McMurray, Aslin, Tanenhaus, Spivey, & Subik, 2008).

In addition to the within-category variation described above, another major source of variation in speech input is due to talker variability. No two talkers sound exactly the same, even if the linguistic content of what they are saying is identical. Abercrombie (1967) coined the term 'indexical' to describe this extra-linguistic variation that exists between talkers: it provides information specific to a particular individual, such as variation related to the properties of the vocal tract or voice quality that might index a talker's gender or age, or idiosyncratic pronunciations

of particular phonemes that would be specific to one particular speaker. Indexical properties can also provide social information about the communities a speaker belongs to, such as pronunciations associated with a particular class or region. Furthermore, indexical information can provide insight into a person's physical or mental state, indexing, for instance, fatigue or affective states such as happiness or anger. All of these indexical properties lead to additional, lexically less relevant variation which speakers must cope with when processing speech. This section aims to provide an overview of relevant research context around the role of talker variability in speech perception: it will first provide an overview of the evidence that shows learners are sensitive to cues which vary between talkers, and that they can learn how these cues are associated with different talkers as well as larger speech communities, before reviewing the evidence concerning effects of talker variability on language processing.

1.1.1 L1 listeners' sensitivity to talker variability

A seminal study investigating the potential effect of talker variation on speech perception comes from Norris, McQueen, & Cutler (2003). They investigated how listeners adapt to speech input from a talker who uses an ambiguous pronunciation (in their case a sound between /s/ and /f/). The key finding was the listeners could cope with this ambiguity, using lexical context to interpret ambiguous sounds as being one or the other category (e.g. where hearing /s/ in 'hou?' would result in *house,* but hearing /f/ would result in the nonword *houf*, listeners were inclined to

interpret it as the sound that would lead to an existing lexical item). They further demonstrated that listeners make use of perceptual learning mechanisms to adapt to the speech input, adjusting their expectations about the talker's pronunciations in the experiment to facilitate resolving ambiguous productions. This line of investigation was continued in Eisner & McQueen (2005), who found that such adjustment is highly talker-specific, and does not generalise to the perception of other novel speakers in the same experiment. This suggests listeners store indexical information about the talker, and learn to use adapted category boundaries for that talker specifically. However, Kraljic & Samuel (2006) showed that for stop contrasts, which contain less information about the talker than fricatives, this learning was not talker-specific and extended to other speakers as well as other phonetic contexts/contrasts (i.e. /b/-/d/ generalised to /k/-/g/). This contrast between fricatives and stops was also found in Kraljic & Samuel (2007) where listeners were presented with multiple speakers with contrasting pronunciations such that one of them used the ambiguous fricative for /s/ while the other used it for /f/. They showed that for fricatives, listeners shifted their category boundaries in a speaker-specific way, suggesting listeners store indexical information about the speaker as well as acoustic information about the contrast itself, while for stop consonants no such speaker-specific adjustment was found but instead the most recent exposure was shown to be used.

However, other work has shown that the extent of generalisation does not only depend on the type of phoneme contrast. Interestingly, people with a smaller social network have been shown to be more likely to generalise idiosyncratic pronunciations to novel speakers when compared to people with a larger social network (Lev-Ari, 2017). The interpretation is that in a smaller social network, each individual has a larger proportional influence, so idiosyncrasies are more likely to be perceived as representative indexical features of variation present within the population at large, while in larger networks, individuals with a deviant pronunciation make up a much smaller proportion of the input a listener receives, facilitating the conclusion that such a deviant pronunciation is idiosyncratic to the speaker. There is also evidence that listeners can use this perceptual learning mechanism for learning systematic phonological processes rather than just phonetic idiosyncrasies (e.g. see Weatherholtz (2015) for learners acquiring vowel chain shifts where a speaker consistently shifted a set of back vowels to be lowered, e.g. /u/ becomes [ʊ], while /ʊ/ becomes [o], and so on). Finally, Kraljic & Samuel (2005) investigated ways in which adjustment caused by perceptual learning could be undone: they showed that time alone is not enough for boundaries to be 'reset'. If a listener hears corrected versions of the token from the same speaker they can unlearn the adjusted category boundaries, but if they hear a different speaker, or the speech from the same speaker but without corrected versions of the shifted tokens, then they do not show evidence of unlearning their adjustments. This

suggests that in order to undo perceptual learning, listeners need to receive evidence that their original source was unreliable.

All in all, there is a consensus in speech perception research that listeners can use perceptual learning mechanisms to deal with variation in the input. Moreover, they do not just "cope" with or discard this variation, but they can actually learn how it is associated with particular talkers, detecting consistent shifts in the distribution of specific contrasts in their input, and evaluating how widely to generalise these shifts based on both the type of contrast (e.g. stops versus fricatives), as well as their linguistic experience (variety of talkers they have been previously exposed to). For an extensive review of further research investigating the role of talker variability in speech processing, see Creel & Bregman (2011).

A further literature explores the factors affecting how listeners learn to associate cues with different talkers using voice or talker identification tasks. Nygaard, Sommers, & Pisoni (1994) demonstrated that participants could learn to identify individual talkers from brief multiple-talker exposure in the laboratory. Further studies have shown that this identification process is aided when the input contains consistent indexical information for specific talkers, e.g. when participants hear stimuli in which each talker has their own characteristic VOT value, compared with stimuli with matched talker variability but no structured phonetic variation (Ganugapati & Theodore, 2018). Voice identification has also been shown to benefit from exposure to variability in terms of speaking styles. Lavan, Knight, Hazan, & McGettigan (2019) demonstrated that voice identification was aided

when stimuli came from multiple speakers who used varied speaking styles (spontaneous casual speech, 3 types of spontaneous clear speech, and read clear speech) compared with when they came from the same set of speakers using a consistent speaking style (read clear speech only). Interestingly, this study found that the benefit of using varied materials (i.e. HV input) specifically held for test items which required generalisation beyond the trained materials, while those trained with one speech style (i.e. LV input) did better for test items similar to the trained stimuli. Note that this specific benefit of variability for generalisation is in line with the results of the phonetic training studies reported in Section 1.3.5.

A final line of research investigating the learning of talker-specific cues has considered how this extends to learning variation that is shared across a group of talkers or a speech community, such as gender- or accent-based variation. Focussing on accents, experiments looking at adaptation to and learning of accents find that this does not seem to occur from the same types of short exposure as picking up on talker-level cues. Shaw et al. (2018) exposed Australian English listeners to a short passage read by multiple speakers of the same unfamiliar regional accent, and then had them categorise nonce words into vowel categories. Testing across multiple different exposure accents, they found that this short exposure to native but unfamiliar accents did not improve performance on the vowel categorisation task in that accent, compared to when the categorisation was performed without exposure to the unfamiliar accent. Similarly, Wade, Jongman, & Sereno (2007) showed that exposing listeners to more varied speech through

means of a categorisation task with feedback did not improve their vowel identification of speech that acoustically varied to the same extent as non-native accented speech would. Further, Kraljic, Brennan, & Samuel (2008) exposed listeners to variation that was either dialectal (i.e. a pronunciation of /s/ as [sʃ] in /str/ clusters only), or idiosyncratic (i.e. a pronunciation of /s/ as [sʃ] across all contexts), before measuring any change to their /s/-/ʃ/ boundary. They showed that even though these cues are acoustically identical, perceptual learning of indexical cues at the accent-level did not seem to occur as people did not adjust their category boundary, while this did occur when the cues were talker-specific.

However, one study did find evidence of dialect-level learning and categorisation, although interestingly only when exposed to multiple talkers in training (Clopper & Pisoni 2004). They used a training paradigm to investigate the effect of talker variability on dialect categorisation, hypothesising that participants who are exposed to more example talkers might be better at being able to categorise novel speakers into dialects at test. Participants were trained to categorise speakers into 6 American English dialect regions. For this they received either high variability (HV, 3 talkers per dialect region) or low variability (LV, 1 talker per dialect region) training input. Results showed that the group who had received LV input did better in the training part of the experiment, as well as in identifying familiar talkers at test, which is evidence of listeners adapting to a specific talker (as seen in previous studies discussed above). However, those listeners who received HV input did

better in the generalisation task. Note that, as with the study by Lavan et al. (2019) concerning variability of talker speaking styles, the benefit of variability for generalization found here is in line with some findings in the L2 phonetic training literature reported in Section 1.3.5. The authors' interpretation is that dialect categorisation requires listeners to generalise dialect-specific cues that hold across different talkers of that dialect. More variable input allows them to create some form of mental representation of the dialect, ignoring less relevant talker-specific cues.

Overall, research on L1 talker variability thus suggests that talker variation and accent variation are processed in separate ways, and while speakers can and do adapt perceptually to speaker-specific variation fairly easily, variation caused by accent or dialect differences seems to be normalised through different processes. However, both research in voice identification as well as accent categorisation finds that having more variable input seems to be helpful in generalisation across talkers.

1.1.2 Effects of variability in L1 spoken word processing

The previous section discussed evidence that listeners can pick up on cues associated with different talkers and accents. These cues are not only relevant in identifying individual talkers or groups of talkers, but may also facilitate language processing more generally. For example, Nygaard et al. (1994) discussed above found that spoken word recognition of novel words in noise was influenced by the

familiarity with the talkers, such that listeners did better if they heard a familiar talker over a novel talker. Further, Creel, Aslin, & Tanenhaus (2008) found that in a lexical identification eye-tracking task in which listeners had to identify words from a minimal pair, they showed less fixation on the competitor items when previous exposure to the foil had been from a speaker with a different gender, compared to from a speaker with the same gender; listeners were thus able to use information about the talker's gender to facilitate their decision.

The processes behind these perceptual learning mechanisms used for speech perception are formalised in the 'ideal adapter' framework proposed by Kleinschmidt & Jaeger (2015), which proposes that listeners continuously use the statistical properties of the input they receive to adapt whatever representation they might have of speech. In this framework, a listener stores indexical information and can therefore recognise familiar talkers without having to relearn statistical properties. Based on those indexical properties, the model can also generalise existing representations to novel talkers, whereas they should adapt to novel patterns that are too dissimilar to the previous experience. This model was investigated further by Kleinschmidt (2019), who posed that to generalise across dialects and genders, listeners will track information for the most informative and useful groupings in the input.

Although learning of structured variation (i.e. non-random variation such as accent variation or indexical variation) can facilitate language processing, there is also a substantial literature demonstrating that encountering multi-talker input can be

detrimental in spoken word recognition. Mullennix, Pisoni, & Martin (1989) investigated the effect of trial-by-trial change in talker on spoken word recognition. Listeners were exposed to stimuli either spoken by a single talker, or by 15 different talkers that varied every trial. They found a consistent effect of talker variability across several different tasks: hearing multi-talker input had a detrimental effect on word identification in noise, word naming (without noise), or identification with degraded signal. A potential explanation for the detrimental effect was provided by Martin, Mullennix, Pisoni, & Summers (1989), who investigated the role of talker variability in spoken word recall. In a set of experiments in which participants learnt lists of words, multiple talker input was found to be detrimental compared with single talker input for the first part of word lists, but this difference disappeared for the later parts of the list. They suggested processing input from multiple talkers takes up more working memory capacity, as the listener has to adjust to a novel speaker on every trial, however this effort is reduced with increased exposure. Interestingly, the effects of variation may be constrained. Sommers & Barcroft (2006) found that variability only has an effect on speech processing when it alters information which the listeners know is potentially relevant for speech perception in their language. This was shown in experiments looking at the effects of variation in speaking style, f0, and speaking rate on spoken word recognition: f0, which is phonetically irrelevant to speech perception in English, did *not* affect English listeners' spoken word recognition,

while phonetically relevant variation occurring due to changes in speaking style and speaking rate showed a detrimental effect on listeners' performance.

Taken together, the studies discussed in Section 1.1.1 and 1.1.2 demonstrate that L1 listeners are sensitive to cues at the level of individual talkers as well as at the level of larger groups or speech communities. They can learn how those cues are associated with individuals and groups of talkers, and can use that information to aid speech processing. Nevertheless, L1 listeners may show a processing cost when encountering varying cues from different talkers, particularly before learners have had a chance to adapt to the talkers in question.

1.2 Second language speech perception

In the first few months of life, infants are able to discriminate most sounds of the world's languages, but over time this ability declines and becomes more specific to those phonemes relevant to the native language or languages the child hears around them (Werker, 1995; Werker & Curtin, 2005). This means that by adulthood, the discrimination of non-native phonemes is often much more difficult.

A key theoretical account of this process is Kuhl's Native Language Magnet theory (NLM, Kuhl, 1993; Kuhl et al., 2008), which proposes infants' perception of their native phoneme categories is warped by their native language experience. With more experience of instances of their native language phoneme categories,

prototypical examples of these native categories start acting like a 'perceptual magnet' for members of the same phoneme category. This creates a distorting effect which ultimately leads to a decreased ability to discriminate 'irrelevant' non-native phonemes, while tuning in to the L1 contrasts. While the NLM theory purely describes infants' speech perception, a slightly different account is provided by the PRIMIR model (Processing Rich Information from Multi-dimensional Interactive Representations, Werker & Curtin (2005)). This framework was developed to be able to explain both speech perception and word learning, and emphasises the role of evolutionary biases for certain aspects of language (e.g. child directed speech, point vowels) in filtering the input over which learning may occur. In addition, it emphasizes that as infants pick up on regularities in the ambient language input, they form multidimensional representations that contain not only phonetic information, but also word-level and indexical information (which we saw in Section 1.1 plays a role in speech perception).

Once a listener's perception has been attuned to their native language(s), any L2 will be perceived through the filter of the L1 experience. A key model of second language speech learning is the Speech Learning Model (Flege, 1995). Flege proposes that mechanisms used to acquire the phonetics of the L1 remain intact throughout life, and can thus be applied to L2 learning. Crucially, native and non-native phonetic categories exist in the same phonological space, leading to potential overlap. If a non-native sound is sufficiently different from the existing native sounds and listeners perceive it to be a different sound, a novel phonetic category

can be created. However, if the non-native sound is close to an existing native category, a novel category might not be created, meaning it will be harder to learn the non-native sound. SLM proposes two mechanisms through which the L1 and L2 interact: category assimilation and category dissimilation (Flege, 2002). Category assimilation occurs when the L2 sound maps onto an existing native category rather than creating a novel phoneme category, while category dissimilation occurs when a novel category is created and the learner adjusts their vowel space to maintain suitable distance between the existing categories. Crucially, as Flege's model describes L2 speech perception in a learning context, these categories can develop as a result of sustained L2 input, so that novel categories can be created later in the learning process, even though a learner might not initially have distinguished the non-native phoneme from their native phoneme categories.

The Perceptual Assimilation Model (PAM, Best (1994, 1995)) provides an alternative account of L2 speech perception. There are three key differences between PAM and SLM: the first is that SLM focusses on speech perception in L2 learning while PAM originally focussed on the perceptual processes for naïve listeners rather than in a learning context. It was expanded to L2 learning in PAM-L2 (Best & Tyler, 2007), where they predict the likelihood of learners being able to distinguish a non-native contrast on the basis of how similar the articulatory settings of the L2 were to those of the native language. This leads into the second difference: while SLM approaches L2 perception from a purely comparative

perceptual point of view where both L1 and L2 phonemes exist in the same phonological space, a key assumption of PAM is that perception of non-native phonemes happens in terms of the similarities and differences to the articulatory gestures of their native phonemes. Thirdly, the models differ in the account of how vowels are represented in the phonological space. Details of this are not relevant to the current thesis, but PAM does provide a more specific account of the ways in which assimilation can occur. It poses that listeners pick up on the discrepancies and similarities between non-native phonemes and their native phonemes. It suggests that assimilation is not an all-or-nothing process, and describes different types of assimilation that may occur depending on the relationship between the non-native contrast in question and the phonology of the L1. It proposes four specific patterns in which these non-native sounds could assimilate to native phoneme categories: *Two Category* assimilation, *Single Category* assimilation, *Category Goodness* assimilation, and *Nonassimilable* contrasts. For *Two Category* assimilation, a non-native contrast is assimilated onto two native categories, resulting in fairly accurate discrimination performance. For *Single Category* assimilation, the non-native phonemes of a contrast map onto the same native phoneme as equally good versions of that category, resulting in conflation of the sounds and difficulties perceiving the contrast. For *Category Goodness*, both sounds of a non-native contrast also map onto a single native category, but they differ in the goodness of fit for that category. This should lead to better discriminability than *Single Category* assimilation. Finally, *Nonassimilable* sounds

are too distinct from any native phonemes to be assimilated into a native category. The discriminability of these sounds will depend on the auditory qualities rather than their likeness to native phonemes. These proposed patterns are borne out in experimental results with adults (Best, McRoberts, & Goodell, 2001), although they have not been tested specifically in children.

These models can account for many behavioural results from adult L2 speakers. Iverson & Evans (2007) looked at perceptual mapping of English vowels by German, Norwegian, Spanish, and French learners. German and Norwegian have a large vowel system, while Spanish and French both have relatively small vowel systems. They found that German and Norwegian learners performed better than Spanish and French learners, as these learners were able to map individual L1 vowels to the English L2 vowels, while Spanish and French learners mapped multiple L2 vowels to one L1 vowel. In terms of the models, this means Spanish and French learners showed single category assimilation, while German and Norwegian learners had two category assimilation (PAM), or might even have showed vowel dissimilation (SLM). Similar research on consonants was done by Iverson, Ekanayake, Hamann, Sennema, & Evans (2008), who investigated L1 category assimilation of the English /w/-/v/ contrast for Sinhala, German, and Dutch learners. Sinhala and German both have just one phoneme that is similar to both English /w/ and /v/, and showed evidence of single category assimilation, while Dutch has two similar phonemes to map onto, and showed a much better

performance in line with two category assimilation. Broersma & Cutler (2011) explore how these effects of L1 perception affect L2 word recognition. Their study looks at phantom activations, where minimal pairs of a difficult vowel contrast become conflated and result in a near-word being activated in perception (e.g. hearing English 'bank' might activate the near-word 'benk' to Dutch listeners who have trouble with the /æ/-/e/ contrast). They found that Dutch listeners were more inclined to judge near-words to be real words than native English listeners, that priming with near-words facilitated word recognition for the Dutch but not English listeners, and that this priming even worked when the near-word spread across word boundaries (e.g. 'evi*l emp*ire' still facilitated recognition of 'lamp' for Dutch listeners). This suggests that the non-native listeners assimilated both vowels in the /æ/-/e/ contrast onto the same category, suggestive of assimilation in SLM, and single category assimilation in PAM (although there may be a category goodness difference). Finally, Escudero & Boersma (2002) found that multi-category assimilation can also occur, such that a binary contrast is perceived as being three or more categories. This was found in Dutch learners of Spanish, who used three Dutch categories to map a two-way front vowel contrast of Spanish, although the effect was reduced if they were aware that they were listening to Spanish (rather than Dutch) stimuli, particularly for advanced learners.

While acknowledging that differences in L1 play a role in L2 perception, experimental research has also looked at performance differences within learners

of the same L1. Holliday (2016) shows that having little experience with the L2 might be more detrimental than being a naïve listener: novice Mandarin learners of Korean showed worse discrimination on the Korean /s^h/-/s*/ contrast than naïve Mandarin listeners did, while performance was similar for naïve listeners and advanced learners. The novice learners have some knowledge of the L2 lexicon, orthography, and phoneme inventory, so the PAM-L2 extension (Best & Tyler, 2007) predicts this influences their perception. Along the same line, Díaz, Mitterer, Broersma, Escera, & Sebastián-Gallés (2016) investigated what might be behind a difference in discrimination ability in a seemingly homogeneous group of advanced learners. They used MMN (mismatch negativity), an event-related brain potential that is elicited upon hearing a deviance from a pattern (Näätänen, 2001), as a measure of perception in a task where listeners had to respond on hearing two instances of the same stimuli in a sequence of alternating ones; they expected participants who had previously shown better L2 discrimination ability to show a bigger MMN than those who showed poorer performance. Interestingly, this bigger response was found in better discriminators regardless of the language the stimuli were in (L1, L2, or a language unfamiliar to the listener), suggesting L2 discrimination abilities are not just a result of perceptual assimilation based on the L1, but also rely on more general speech processing abilities.

Although the models described above are not directly tested in this thesis, they provide a background for interpreting the results. Crucially, this thesis uses phonetic training to improve perception of L2 vowel contrasts, focussing on vowel

contrasts in the L2 that do not exist in the participants' L1 and which are known to lead to difficulty in perception (see Chapter 2). It is therefore worth considering what might be expected of the non-native participants' performance when using these contrasts in identification and discrimination tasks administered pre- and post-training. Importantly, for the current thesis, these models are not mutually exclusive in their predictions. Both predict assimilation of the non-native phonemes onto existing native categories: PAM just makes more specific predictions than SLM does, while SLM considers what might change as a result of increased exposure in a learning context.

In terms of PAM, the use of contrasts that are notoriously difficult for these participants means there will be a tendency for these contrasting phonemes to be assimilated into a single category where identification and discrimination are both similarly difficult. However, the model also allows for there to be single category assimilation with a category goodness difference, where one of the two phonemes is a better fit to the category than the other. This might make them easier to discriminate than identify: discriminating them can be done on the basis of how good of an exemplar of the assimilated L1 category they are, while identifying them relies on the ability to group both exemplars as belonging to different L2 categories despite having assimilated them to the same L1 category. This effect of assimilation with a category goodness difference has specific predictions for participants' performance: discrimination performance is expected to be better than identification in general. The choice to work with notoriously difficult contrasts

would mean SLM also predicts all contrasts to undergo category assimilation, where the two vowels map onto one existing L1 vowel without creating a new category. Further exposure and training might lead to category dissimilation and the creation of a new category so that performance should improve after exposure to the training materials (pre- to post-test improvements).

1.3 Second language phonetic training

The preceding sections have shown that learning non-native speech sounds can be difficult. This leads to the question of to what extent they can be learnt, and whether providing any training is useful. This has been looked at in the phonetic training literature. The foundations of phonetic training were established in the 1970s and 1980s. One of the more influential initial attempts at training learners on a specific phonetic contrast was by Strange & Dittmann (1984), who used a continuum of synthesised speech in phonetic training to train Japanese learners of English on the /l/-/r/ contrast using an AX-discrimination task with feedback in training, and aimed to see if this transferred to natural speech at test. Learners showed a significant improvement in their /l/-/r/ discrimination accuracy and their identification ability, but did not generalise to natural speech. Jamieson & Morosan (1986) posited that this lack of transfer to natural speech was down to three factors: stimuli being presented without an appropriate acoustic context, the absence of acoustic uncertainty (i.e. variability) in the stimuli, and the use of a discrimination

rather than an identification task in training. In their follow-up study, Jamieson & Morosan (1986) aimed to show that once these principles were taken into account, successful transfer from synthesised speech to natural speech would occur. They presented francophone Canadian adults with synthesised English /ð/ or /θ/ tokens using more varied phonetic contexts (i.e. a set of CV syllables) and with stimuli that also varied along a continuum between prototypes of the two phonemes. They presented these stimuli to participants using a fading task, in which learners had to first identify tokens from the continuum endpoints before gradually introducing more instances closer to the category boundary. They found that learners could generalise to natural speech following this training. Jamieson & Morosan (1989) found that the use of an identification task without the varied stimuli but using only prototype tokens also resulted in some generalisation to natural voices, but less so than when using the variable stimuli and fading task. Together, these studies suggest that both using an identification task, and including variable stimuli aid generalisation from trained synthetic voices to natural voices.

A further turning point in the phonetic training literature occurred in the early 1990s, when researchers began using natural speech in phonetic training. Logan, Lively, & Pisoni (1991) again focussed on Japanese learners of English acquiring /l/-/r/, but in contrast to previous training studies used natural speech where the contrasts occurred in multiple contexts and were spoken by multiple speakers. Training took place over multiple sessions and consisted of an identification task with trial-by-trial feedback on their performance; learning was assessed using a

pre-test/post-test design, i.e. where participants are given the same battery of tests before and after training so that improvement can be measured, with a generalisation task afterwards. They found that using this paradigm, participants improved on trained stimuli as well as untrained stimuli using both novel voices as well as novel items. This initial work was further expanded on by Lively, Logan, & Pisoni (1993), who examined the effect of different types of variability in the input stimuli. Their key finding was that participants' ability to generalise to novel stimuli produced by a new talker was specifically dependent on exposure to training stimuli spoken by multiple talkers, while the manipulation of phonetic environment did not influence improvement. Further evidence in support of this came from Magnuson, Yamada, Tohkura, & Bradlow (1995) who again trained Japanese learners of English with LV input, using a slightly larger sample of 10 participants per condition. They again found that single-talker training allowed for generalisation to novel items, but mostly not to novel talkers.

Further work explored the robustness of the *high variability phonetic training* method (HVPT). Lively, Pisoni, Yamada, Tohkura, & Yamada (1994) found that learners were able to fully retain their trained skills at a three-month follow-up, and though a six-month follow-up showed some decreased performance, learners were well above their initial pre-test level. Moreover, Bradlow, Akahane-Yamada, Pisoni, & Tohkura (1997) found that perception training substantially transferred to production of the contrast as well, and Bradlow, Akahane-Yamada, Pisoni, & Tohkura (1999) extended this by investigating the long-term retention of the learnt

contrast in both perception and production, finding that trained as well as generalisation skills were retained at a three month follow-up for both perception and production. This set of findings solidified the case for the use of high variability input, including the use of multiple talkers, in phonetic training, and this became a standard methodology in the field (see for instance (Fuhrmeister & Myers, 2017; Hazan, Sennema, Iba, & Faulkner, 2005; Iverson et al., 2001; Iverson & Preece-Pinet, 2008; Ylinen et al., 2010)).

Theoretically (and intuitively), it makes sense to see an advantage of high variability input on generalisation tasks. Encountering variability helps the listener to recognise which acoustic cues are irrelevant, and thus to focus on those cues that are key to distinguish the phonemes they are being trained on. For example, a speaker from London will tend to sound quite different to a speaker from Yorkshire, and a 7-year-old sounds different from a 70-year-old. If you only ever heard L2 phonemes produced by one talker, you might think that aspects of that speaker's pronunciation are relevant to distinguishing the phonemes, when they are actually idiosyncrasies of the speaker. This could make it difficult to adjust to novel speakers who say things differently from the talker you have heard. In contrast, when you are exposed to multiple speakers who all pronounce the phonemes slightly differently, you can learn which cues are irrelevant and variable (e.g. ruling out pitch as a phonetically relevant cue in SSBE after hearing speakers of multiple ages and genders), and instead focus on those cues that remain stable throughout (e.g. formant frequencies for SSBE vowel quality). This is in line with

the predictions of computational models where irrelevant contextual or speaker identity cues compete with phonetically relevant cues, meaning that dissociation of the irrelevant cues is key to generalisation (Apfelbaum & McMurray, 2011; Ramscar, Yarlett, Dye, Denny, & Thorpe, 2010).

Before going into more detail, an important thing to consider is what we mean when we talk about phonetic learning, as this is essentially the target outcome of phonetic training. The goal of phonetic training is often to improve the perception and/or production of a particular phoneme contrast. It might in the first instance result in a better perception ability where learners are able to identify a phoneme they were not able to accurately identify before. This perceptual improvement resulting from phonetic training has been found to generalise to novel stimuli and talkers, and even to larger discourse contexts (Huensch, 2016). However, there is a limit to this generalisation: the improvement after vowel training does not seem to extend to vowels that have not been trained (Nishi & Kewley-Port, 2007), nor does consonant training necessarily seem to generalise to these same consonants in different phonetic contexts (Shinohara & Iverson, 2013). The improvement may be task specific, where it is seen in identification but not discrimination (Gong, Yang, Ji, & Wang, 2019). Short-term training often also does not seem to result in much generalisation at all (Logan & Pruitt, 1995). Moreover, while there is evidence of retention up to 3 months (Bradlow et al., 1999) and even 6 months (Wang, Spence, Jongman, & Sereno, 1999), exactly what and how much is retained seems to vary. Whether the results of phonetic training extend to changes in the underlying

representations remains to be seen: some literature suggests phonetic training studies can result in improvement of perceptual abilities, but that it does not necessarily affect the phoneme categories. (Iverson & Evans, 2009) who found that while phonetic training did improve adult learners' ability to identify the non-native phonemes, they did not show any change to their best exemplars. Similarly, (Heeren & Schouten, 2010) found improvement in children's identification ability after phonetic training, but saw little change in their category boundaries. This suggests that what phonetic training might be doing is changing the way the perceptual cues present in the speech input learners receive are being processed, or potentially how the cues are being weighted in the processing, rather than adjusting the underlying representations of the phoneme categories.

Section 1.3.5 will return to consider the question of how good the evidence for an HV advantage in phonetic training actually is. First, however, the next few sections will consider the way in which the HVPT methodology has been extended in the literature. In particular, it will consider the fact that these studies each use somewhat different adaptions of HVPT. Four particular aspects are relevant to choices made in this thesis, and they will each be discussed in turn: specifically, the effect of the type of task that is used in training (discussed in Section 1.3.1), the number of contrasts that are being trained (Section 1.3.2), the use of production as well as perception in training and/or at test (Section 1.3.3), and the extension of the paradigm to child learners (Section 1.3.4).

1.3.1 Type of training task

As described above, the initial phonetic training studies used discrimination-based training (Strange & Dittmann, 1984). However, following the work by Jamieson & Morosan (1986) and Lively et al. (1993), HPVT training generally uses identification tasks rather than discrimination tasks in training. This shift is likely due to a recommendation by Logan & Pruitt (1995), who pose that discrimination tasks are generally less suited to train novel phoneme categories, as they focus on detecting any differences between the phonemes. This could make it harder for learners to figure out the similarities across categories, as they have to filter out existing within-category variation. Several studies have continued the work on Japanese learners of English /r/-/l/ showing successful learning and generalisation through means of identification training: Iverson et al. (2001, 2003) looked at /r/-/l/ for Japanese, German and American English adults, to investigate the differences in perceptual sensitivity to the different cues for /r/-/l/ between the language groups, while Shinohara & Iverson (2013, 2015) expanded the /r/-/l/ investigation to training children (further discussed in section 1.3.4). Iverson, Hazan, & Bannister (2005) provided a comparison of different stimuli types used to train Japanese learners of English /r/-/l/, finding that both HVPT with natural voices, as well as variable stimuli created through different types of signal processing were effective in improving learners' performance. Only a handful of studies have

looked at adult perceptual learning of other consonantal contrasts, all using identification training: Sadakata & McQueen (2013) successfully trained Dutch learners on Japanese geminate consonant contrasts, Fuhrmeister & Myers (2017) taught American English speakers the Hindi dental versus retroflex /d̪/-/ɖ/ contrast, Iverson et al. (2008) examined the difference between native Sinhala, German, and Dutch speakers learning the English /w/-/v/ distinction, and Kim & Hazan (2010) trained British English speakers on the Korean /t/-/t^h/ contrast.

However, the role of discrimination in training has hardly been investigated since the 1995 advice to use identification tasks instead. The first and only study to do so is the recent Shinohara & Iverson (2018), who investigated the effect of using identification versus discrimination tasks in training Japanese learners of English /r/-/l/. They investigated the effects this had on pre/post-tests of identification, auditory discrimination, category discrimination, and production. They concluded that both methods proved similarly successful in eliciting improvement across the board, although a slight advantage for identification training was found in the post-test identification task. This advantage was attributed to a task effect, as the identification task at post-test was nearly identical to the training. Overall, their results cast doubt on the original claim that discrimination is not an ideal training task. Nevertheless, the use of identification tasks in training is much better established in HVPT and was thus used in the current thesis.

The studies described above which used an identification task made use of orthography, i.e. participants were trained to identify which of two letter symbols (e.g. 'r' or 'l') or letter sequences (e.g. 'ee' versus 'i') corresponded to the phoneme heard in a given trial. However, it is possible to have an identification task without using the orthography. This has primarily been seen in studies using HVPT to train listeners on the lexical tone system of Mandarin Chinese. These studies are particularly interesting because they often involve naïve learners. This means that, in order to use an identification task in which the learner maps the stimuli to the tone, they need a system for depicting the suprasegmentals, without being able to rely on an existing system of familiar orthography (a problem already raised in Logan & Pruitt (1995)). Several of these tone training studies use numbers to represent the different tones (Sadakata & McQueen, 2014; Wang et al., 1999). Another set uses arrows depicting the direction of the pitch contour (Antoniou & Wong, 2015; Ingvalson, Barr, & Wong, 2013; Perrachione, Lee, Ha, & Wong, 2011), while Wang & Kuhl (2003) used pictures of animals to be associated with individual tones by children and adults (see also Section 1.3.4). However, two studies in the literature avoided representing tones directly and instead used an identification task in which participants mapped whole words to pictures: Wong & Perrachione (2007) trained adults to associate meaningful pictures with 18 Mandarin Chinese nonce words, where the words formed minimal pairs only distinguished by their use of lexical tone. They found significant learning and generalisation to new talkers at test. Dong, Clayards, Brown, & Wonnacott (2019) also used a picture-

mapping task, here using real Mandarin words and associated pictures, and again found evidence of learning and generalisation. Outside of tone learning, a picture-mapping training task has also been used in two phonetic training studies with child learners acquiring non-native vowel contrasts (Evans & Martín-Alvarez, 2016; Giannakopoulou, Brown, Clayards, & Wonnacott, 2017); these are described in further detail in Sections 1.3.3 and 1.3.4.

One question that remains is whether the use of orthography, or orthographic type symbols, is helpful in training. As discussed above, most training studies with adults have made use of these symbols, and there is some evidence that they are useful for these learners. Escudero, Hayes-Harb, & Mitterer (2008) taught Dutch adult learners of English /e/-/æ/ minimal pair words in a picture-matching task, where groups were either trained with auditory information only, or with auditory and orthographic information. Then, in an eye-tracking task probing word recognition, learners who had been exposed to orthography in training were more consistent in looking at the target for the /e/-words, while those who were only presented with auditory information looked more at distractor items for both vowels. The authors interpreted this as evidence for a beneficial role of orthography in training non-native phonemes. On the other hand, Simon, Chambless, & Kickhöfel Alves (2010) found no added benefit of orthography. They investigated the acquisition of a French vowel contrast in American English learners, and found that the addition of orthographic information over auditory information made no difference to the

performance on a picture-matching word learning task or an AXB discrimination task. However, note that both participant groups showed near-ceiling performance, potentially masking any effect of orthography. The role of orthography in the acquisition of a non-native phoneme contrast for children remains an open question, as there is no research specifically investigating this to date.

The current thesis makes use of identification tasks in all studies. Study 1, where participants were naïve learners, investigated both mapping the novel phonemes to orthography-like symbols, as well as to whole-picture mapping.

1.3.2 Number of trained contrasts

While studies training consonants and tones have increased in number over the years, the majority of studies using HVPT have investigated the learning of non-native vowels. With a few exceptions, by far the majority of these have worked with second language learners of English, reflecting the current English-dominant attitude in academic research as well as in second or foreign language teaching. In general, vowel training studies have targeted single vowel contrasts that were notoriously hard for the learner group in question: several studies investigated training of the English /i/-/ɪ/ contrast for Finnish learners (Ylinen et al., 2010), Spanish learners (Evans & Martín-Alvarez, 2016), Greek learners (Giannakopoulou, Uther, & Ylinen, 2013), and Japanese learners (Grenon, Kubota, & Sheppard, 2019). Similarly Wong (2014) trained Cantonese speakers on English

/e/-/æ/, additionally investigating the role of proficiency in the success of learning in a follow-up study (Wong, 2015b). For learners of non-English vowels, Kartushina & Martin (2019) investigated Spanish learners of the French /e/–/ɛ/ contrast through means of articulatory HVPT training, Kartushina, Hervais-Adelman, Frauenfelder, & Golestani (2015) trained French speakers on the Danish /e/-/ɛ/ and /y/-/ø/ with production training to explore the link between perception and production, and Alispahic, Escudero, & Mulak (2014) used HVPT to teach the Dutch /ɪ/-/ʏ/ contrast to Australian English and Peruvian Spanish speakers.

However, there is evidence that training on a single vowel contrast might not be ideal in teaching learners non-native vowels. Nishi & Kewley-Port (2007) investigated the effect of using a subset of the vowel system versus the full vowel set with Japanese learners of American English vowels. In the subset group, learners were trained on 3 difficult vowels (/ɑː, ʌ, ʊ/), while the fullset group received training on all 9 monophthongs, covering the entire vowel system. They found that training was less successful when training on the subset only. Overall performance was lower in the subset training where no improvement was seen on the untrained vowels, and where performance did not seem to generalise to novel words and was retained for a shorter period of time. The authors subsequently extended this to investigate whether a hybrid fullset-subset training was equally effective as fullset training. They found this to be the case, as long as the hybrid training started with the fullset and then focussed in on the subset. They

interpreted this as showing that if training first focussed on a vowel subset, learners initially created novel categories for the subset, causing difficulty when these needed to be adjusted to accommodate the rest of the vowel set (Nishi & Kewley-Port, 2008). This suggests learners might benefit from larger vowel sets in training, rather than selected vowel contrasts.

While some studies have still used a subset of contrasts known to lead to difficulty for the learner group (e.g. Rato (2014); Rato & Rauber (2015) for Portuguese learners, Aliaga-García (2009); Aliaga-García, Mora, & Cerviño-Povedano (2011); Carlet (2017); Carlet & Cebrian (2015) for Catalan learners, and Wong (2015a) for Cantonese learners), many further studies have since successfully used the full English vowel set to train various groups of second language learners (see Lengeris & Hazan (2010) for Greek learners, Lengeris (2009) for Greek and Japanese learners, Højen & Flege (2006) for Spanish learners, Iverson & Evans (2009) for German and Spanish listeners, Iverson, Pinet, & Evans (2012); Iverson & Preece-Pinet (2008) for French learners, Thomson (2011) for Mandarin learners). There are also a few studies which combine the training of (in this case English) vowels and consonants, either restricting it to those phonemes that are particularly difficult for the non-native speakers (Cebrian & Carlet (2014) training Catalan and Spanish learners on English /b/-/v/, /d/-/ð/, /æ/-/ʌ/, and /i/-/ɪ/), or training them on a very large set of both vowels and consonants to investigate differences in phoneme acquisition (Hwang & Lee (2015) for Korean learners of English).

Overall, this literature shows it is possible to train on the whole non-native vowel set, and that this might even be advantageous for learning individual contrasts. In the current thesis, both approaches are used: Study 1 follows the majority of the literature in using single-vowel contrasts, while Study 2 looked at the learning of multiple vowel contrasts in parallel.

1.3.3 From perception to production

The majority of studies discussed above have primarily used perceptual training and tested for improvements in perception. An important further question is whether training can also lead to benefits in production. A link between perception and production is well-established in the literature (though the link might be relatively weak, cf. Hattori & Iverson (2009); Mitterer & Ernestus (2008)), and such a link predicts that knowledge acquired through perception could transfer to production. However, experimental results here are mixed (see Sakai & Moorman (2018) for a meta-analysis). As mentioned above, initial studies by Bradlow and colleagues found that perceptual training on the /r/-/l/ contrast transferred to production, and observed that this transfer could be retained for at least three months (Bradlow et al., 1997, 1999). Further links between perception and production were found in Lambacher, Martens, Kakehi, Marasinghe, & Molholt (2005), who trained Japanese learners on American English mid and low vowels, and in Wong (2014, 2015b), who trained Cantonese learners on English vowels. A positive transfer from perception training to production has also been found for

production of non-native syllable structures (Huensch & Tremblay, 2015), and for contrasts embedded in larger discourse contexts (Huensch, 2016).

However, several studies have also shown that improvement in perception does not predict production or vice versa: while Iverson et al. (2012) found native French speakers improved in their perception and production of SSBE vowels, they found no evidence of a correlation between participants' improvement in either domain, and Shinohara & Iverson (2018) find a similar lack of a link between the amount Japanese learners improved on SSBE /l/-/r/ in perception and production.

More mixed evidence comes from Alshangiti (2015); Alshangiti & Evans (2014), who taught Arabic learners of English on SSBE vowels, finding that learners only improved on production if their training included specific production training as well, while perceptual training alone did not transfer to production. Relatedly, Hwang & Lee (2015) found very little change in either vowel or consonant production in Korean learners of American English as a result of perception training. However, it is important to be cautious interpreting null findings as it could be the case that the data is underpowered, and thus not sufficiently sensitive to test the hypothesis (see Section 1.5 for more discussion of this general problem in the literature). Keeping this in mind, the research does tentatively suggest that perceptual training seems to lead to improvements in production, though this may not always be the case. The current thesis uses perceptual training only in both studies, but in Study 2 a production task is included at test to see if there is any evidence of transfer from perception to production.

1.3.4 Training children

Most studies using HVPT have been with adult learners, although more recently studies have also been done with children. The key question of interest in this work has generally been whether younger learners will show greater improvements than older learners in acquiring novel L2 contrasts, following the Critical Period Hypothesis (Lenneberg, 1967) discussed at the start of Chapter 1. The first to investigate this in the context of HVPT was the study by Wang & Kuhl (2003) (introduced in Section 1.3.1) who trained 6, 10, and 14-year-old American children as well as adults on Mandarin tones using 6-session HVPT using an identification training task in which participants learned to associate tones with pictures of animals. Overall, adults outperformed children and older children outperformed younger children at both pre- and post-tests. However, all age groups improved their tone identification after training in similar proportions, suggesting no evidence of a difference in learning rate for the different age groups. Giannakopoulou et al. (2017) also investigated whether there was an age difference in the ability to benefit from phonetic training, comparing Greek 8-year-olds and adults learning the English /iː/-/ɪ/ vowel contrast after 10 HV or LV training sessions. While the study was primarily about phonetic training, it also included tests probing vocabulary learning (discussed further in Section 1.4). Their findings were in line with those of Wang & Kuhl (2003), with adults starting off with a higher performance, but both age groups showing similar amounts of learning on identification tasks. Similar results were found by Heeren & Schouten (2010), who

investigated Dutch adult and 12-year-old child learners' ability to acquire the Finnish /t/-/tː/ length contrast using HVPT identification training in which stimuli had to be labelled as 'short t' or 'long t'. They found that although adults started off with a higher performance, both adults and children showed similar improvement on identification and discrimination tasks after 3 training sessions.

Some other studies *have* found age-based differences in terms of learning rate. Giannakopoulou, Uther, & Ylinen (2010); Giannakopoulou et al. (2013) tested Greek adults and 7-8 year old children learning the English /iː/-/ɪ/ contrast after 10 sessions of HVPT mapping to orthography. While both adults and children improved their discrimination and identification abilities after training, children showed greater improvement than adults. However, it is worth noting that adults started out with higher performance in the relevant pre-tests, meaning that children had greater room for improvement. Age differences were also found in Shinohara (2014); Shinohara & Iverson (2013), who investigated the effect of training in four different age groups: young children (6-8), older children (8-12), adolescents (15-18), and adults. All were Japanese learners trained on the English /r/-/l/ contrast using 10 HVPT sessions. In line with the prediction that younger learners show more plasticity while adults show more entrenchment of L1 phoneme categories, adolescents and older children improved their identification accuracy and discrimination ability more than the adults did. Interestingly, however, the youngest group of 7-8 year-olds did *not* outperform adults, and in

fact did not improve their performance after training at all. The authors attributed the lack of improvement to the children's phonemic awareness still being in development. Phonemic awareness is a term used in the literature on literacy development as part of a larger set of phonological awareness skills, and is defined as the ability to break words down into individual phonemes. This is an important stepping stone for the development of L1 literacy skills, where letter-sound mappings need to be learnt (Snowling & Hulme, 1994). This link to orthography is key to the explanation Shinohara (2014); Shinohara & Iverson (2013) provide for the poor performance in their youngest group: their study used orthography in training to depict the two phonemes. If the children have difficulty hearing the different sounds within the words, and if they are less familiar with the concept that letters can consistently map to those sounds, they might not do well with orthography-based training. It is worth noting that though the children in Giannakopoulou et al. (2010, 2013) and Heeren & Schouten (2010) also mapped to (a form of) orthography, the orthography in question had a length distinction (i.e. 'ee' versus 'i' in Greek and 'short t' versus 'long t' for Finnish). This may have made it easier to map the orthography to the longer sound, via iconicity. Children may find it more difficult to map sounds to symbols when the orthographic mapping is more arbitrary.

Overall, there is no good evidence that children learn more in these perceptual training paradigms than adults. However, in all studies described above, apart from Wang & Kuhl (2003), adults start out with more experience of the target language,

potentially facilitating further learning. It is important to further investigate what happens when both adults and children start out with the same amount of experience. To do this, Study 1 of the current thesis uses naïve adult and child participants who have no knowledge of the target language. The literature reviewed here also demonstrates that certain tasks might be more difficult for children than adults (e.g. mapping to orthography). The current thesis sheds further light on the types of tasks that work with learners of different ages.

1.3.5 The role of input variability in phonetic training

Given these mixed results with regard to an age benefit for children discussed above, an important question is whether the HVPT training material is well suited for younger learners. In particular, is *variability* in the input actually beneficial? This is the key question in the current thesis. However, before considering the relevant literature with children, it is worth noting that, even for adults, while numerous studies have used HVPT, few have directly compared the efficacy of HV versus LV input. Returning to those seminal studies, Jamieson & Morosan (1986) and Morosan & Jamieson (1989) found a kind of variability benefit in that they got more generalisation in the experiment with stimuli that varied along a continuum than with non-variable prototype stimuli. However, the sample sizes are rather small (with 10 and 12 people in the treatment group, respectively), the different methods across the studies make it hard to draw direct conclusions about the role of variability per se, and note that this is synthetic variability rather than natural talker-based variability. Returning to the next set of seminal studies, note that

Lively et al. (1993) and Logan et al. (1991) never made a direct comparison between the amount of improvement for the LV study and the HV study. The HV advantage they observed was based on observations of two separate experiments (reported across the two publications), and they never statistically compared the amount of improvement found in each of the variability conditions. Additionally, while 6 participants were trained in Logan et al. (1991), generalisation to novel items was only tested in 3 participants.

More appropriately powered direct comparisons of variability input in phonetic training have been done more recently, but are relatively few in number. Recall that two example studies where a variability manipulation affected generalisation were already discussed above. Lavan et al. (2019) found a benefit of hearing varied speaking styles in a voice identification task, and Clopper & Pisoni (2004) found a benefit of hearing multiple voices in a dialect identification task. In both cases, a high variability benefit was seen in generalisation tasks, in line with the explanation that variability helps participants to ignore irrelevant cues and focus on the stable cues which are relevant to the tasks.

Returning to second language phonetic training, Kartushina & Martin (2019) used production training with either HV or LV talker input to train adult Spanish learners on a French vowel contrast. After 3 training sessions, while production accuracy improved for both conditions, only HV participants showed generalisation to items spoken in a novel voice. Additionally, the HV participants were more stable in their productions than those receiving LV input. For

perceptual training of non-native segmental contrasts, three further adult studies have directly contrasted HV and LV input in phonetic training. The first was a training study by Sadakata & McQueen (2013) in which Dutch adult participants were trained to acquire a Japanese geminate consonant contrast in 5 training sessions of either HV or LV training input. Participants showed better learning as well as generalisation in an identification task after HV training than LV training, although their discrimination performance improved regardless of training variability. Similarly, Wong (2012) trained Cantonese learners on English vowel contrasts for 10 sessions of HV or LV input. Both groups improved after training, though those with HV input outperformed those with LV input and also showed more generalisation on identification tasks; this result was replicated with learners of high and low proficiency levels in Wong (2014) (thus speaking against the case for an interaction).

The role of variability has also been investigated in phonetic training of lexical tones. In an eight session training study, Perrachione et al. (2011) found that when there was trial-by-trial variability, HV was beneficial specifically for learners who had stronger perceptual abilities (as measured on a pitch-contour perception test), while learners with weaker perceptual abilities suffered from having HV. Sadakata & McQueen (2014) came to a similar conclusion in their 5 session tone-training: learners who had a low perceptual aptitude were hindered by having more variability, while those with a high aptitude benefitted from it. However, a similar eight session tone training study for English learners of Mandarin tones by Dong

et al. (2019) did not find either an overall benefit of training with HV materials, nor an interaction with individual aptitude. This study also used Bayes Factors to investigate whether there was substantial evidence for the null, which they found to be the case for an overall HV benefit. However, for the interaction with aptitude the evidence was ambiguous, with further analyses suggesting that although their participant sample (60 participants) was larger than in many previous studies, far larger numbers ($N > 300$) would be necessary to find substantial evidence concerning the interaction. This suggests that previous studies in this area may be underpowered.

If differences in ability between adult participants can affect ability to benefit from HV materials, one might ask if this might be affected by age differences, since child learners are known to differ in their cognitive abilities compared with adults. There are just two studies in the literature which have compared high and low variability training in child learners. Evans & Martín-Alvarez (2016) trained 44 Spanish children aged 9-12 years old on an English vowel contrast through means of either HV (4 talkers) or LV (1 talker) perceptual training input using a whole-picture identification task, aiming to improve perception on a discrimination task as well as investigate any transfer to production. Children trained with LV input improved more over the 5 training sessions than those trained with HV input, and only children trained on LV improved on production, while HV input was key for improvement of vowel discrimination. In terms of the expected HV benefit, no such overall benefit was found even though all test items involved a novel talker

and thus a type of generalisation. However, an HV benefit was seen when also extending to novel items, suggesting that the benefit might only be seen when stimuli are further removed from the training set. The second study comparing variability conditions is the one by Giannakopoulou et al. (2017), introduced in Section 1.3.4. This study used the same training stimuli and methods as Evans & Martín-Alvarez (2016), except it used more training sessions and did not test production. Unlike the Evans & Martín-Alvarez study, this study did *not* find an HV benefit in the discrimination test, either for child or adult learners. It thus also contrasts with the results of the studies which did find an HV advantage in adult generalisation described above. In fact, in the discrimination task, while both groups improved their discrimination ability from pre- to post-test, adults improved the same amount in the high and low variability conditions while children actually showed an unexpected LV benefit that even held for generalisation across novel speakers and items. The authors are cautious in interpreting these results. For adults, they note that the lack of high variability result could be due to the fact that performance was near ceiling at post-test, having started with very high performance at pre-test compared with children, due (at least in part) to their greater previous exposure to English. This highlights the difficulty in comparing adults and children when they have very different starting points. It is also important to note that this study used a discrimination rather than an identification task at test, and the results of Sadakata & McQueen (2013) indicated that the high variability benefit may only show up in identification in

adult learners. For children, Giannakopoulou et al. (2017) were again cautious about the unexpected LV benefit, here noting that there were accidental differences between the groups at pre-test. Although these differences should have been controlled for statistically, this could still have affected the results.

Nevertheless, the authors did consider why it might be that children show a benefit in the LV condition. They noted that an analysis of the data from training showed that both age groups showed a benefit of LV over HV input during training. They discuss this in terms of the literature discussed in Section 1.1.2, which demonstrates that there can be a processing cost when encountering multiple talker input in child language learning, and even for adult listeners in their native language. Giannakopoulou et al. (2017) suggest that high variability material may cause particular difficulty for children, perhaps due to the greater difficulty they may have in adapting to multiple talkers during training due to lower working memory capacity. This processing difficulty could outweigh potential benefits of encountering multiple talker input for generalisation. A relevant detail is that Giannakopoulou et al. (2017) used trial-by-trial talker variation in their input (i.e. talkers were intermixed rather than blocked), which has been shown to be more detrimental compared to blocked variability in learning lexical pitch contours (Dong et al., 2019; Perrachione et al., 2011), as well as in more general areas of speech recognition (Mullennix et al., 1989) and word recall (Martin et al., 1989). Nevertheless, Evans & Martín-Alvarez (2016) also used training materials with

trial-by-trial variability, making it surprising that an HV advantage was seen in that study (at least for novel items).

To the best of my knowledge, these are the only two studies which have compared HV and LV input materials in child phonetic training, and the results are contradictory. There is one other study with child learners which compared the learning of HV and LV materials; this was not in phonetic training, but rather in the related area of vocabulary learning. Since the current thesis also includes tests of vocabulary learning, this literature is reviewed in Section 1.4 below.

1.4 High variability training in vocabulary learning

While most of this thesis is about L2 phonetic learning, some of the tests also tap vocabulary learning. This section reviews the evidence suggesting input variability also plays a role in vocabulary learning. Here, a number of studies have shown that when adult participant are taught and tested on novel vocabulary, they show better recall at test for words they had heard spoken by multiple talkers or speaking styles, compared with those spoken by a single talker or in a single speaking style (Barcroft & Sommers, 2005). This variability benefit also extended to L2 vocabulary recall in noise (Sommers & Barcroft, 2011). The same authors further investigated the type of variability which benefitted learning. Recall from Section 1.1.2 that they also found that encountering variable cues was *not* beneficial but instead detrimental to word recognition when the cues in question were ones which the participants

would expect to be phonetically relevant in their L1 (Sommers & Barcroft, 2006). Here, Sommers & Barcroft (2007) investigated the role of phonetically relevant versus irrelevant variation in L2 vocabulary learning. For English learners, no variability effect was found when investigating phonetically irrelevant variation in amplitude and f0, while for speaking rate, which is phonetically relevant for English, a variability benefit was found. In a follow-up study, Barcroft & Sommers (2014) confirmed that for speakers of Zapotec, where f0 is lexically contrastive and therefore phonetically relevant, variation in f0 was beneficial for L2 vocabulary learning. This fits with an explanation in which encountering variability is beneficial because it provides evidence about which cues are idiosyncratic to the speaker or which cues are irrelevant. For cues where the learner already expects them to be irrelevant on the basis of their L1 knowledge, added variability will not make a difference.

These findings from the adult L2 vocabulary learning literature are notably similar to those found in (L1) infant word learning. This comes from a literature investigating a surprising phenomenon whereby there is a point in development where infants are able to discriminate individual phoneme contrasts but cannot tell apart minimal pairs differing only in those phoneme contrasts (e.g. in Stager & Werker (1997) 14-month-old infants could discriminate /b/ and /d/ but not differentiate between /bɪ/ and /dɪ/ in the context of a word learning experiment; see Werker & Curtin (2005) for a review). Rost & McMurray (2009) proposed that

a lack of input variability might play a role in the difficulty infants have in learning minimal pairs in this task. They investigated this by teaching infants two novel words and contrasting single-talker and matched multiple-talker input. Infants in the first group failed the task (as in previous experiments) whilst those who had received multi-talker input were able to correctly map the minimal pairs to different referents. Infants, like adult L2 learners, thus benefit from variation in indexical properties. Further experimental studies and computational modelling suggest that this benefit arises as in single-talker input, irrelevant talker-specific acoustic features become associated with the words being learnt, while this is not the case for multiple-talker input (Apfelbaum & McMurray, 2011; Rost & McMurray, 2010) or deliberately varied single-talker input (Galle, Apfelbaum, & McMurray, 2015). This explanation is consistent with the idea that variability helps the learners to figure out which cues in their input are relevant to the task.

Cognitively, children can be said to be midway between infants and adults. However, the field investigating the role of variability in child L2 acquisition is surprisingly sparse. To date, there is just one study directly investigating the role of variability in child L2 vocabulary learning. This study corroborated the finding of a benefit of talker variability in L2 vocabulary with adults, but not with children. Sinkevičiūtė, Brown, Brekelmans, & Wonnacott (2019) compared the effect of speaker variability on L2 word learning in an experiment with adults, 7-8 year-olds and 10-11 year-olds. Adults showed the predicted benefit in recall following training with multiple talkers, however, neither of the child groups showed

evidence of a high variability benefit in either tests of production or comprehension. Importantly, this study used Bayes Factor analyses and thus was able to demonstrate that – at least in some of the tests – there was evidence for the null, i.e. for no variability effect (see Section 1.5 for further discussion of Bayes Factors). Interestingly, for the youngest group, there was direct evidence that they had particular difficulty in adapting to the multiple talkers during the exposure part of the study (where a 2-alternative forced-choice task was used). Again, this is in line with the findings discussed in Section 1.1.2, that processing multiple talker input can lead to difficulties even in L1 speech perception depending on the demands of the task. Multiple talker input may thus place a burden on working memory. As children are known to have lower working memory than adults (Alloway, Gathercole, & Pickering, 2006; Case, Kurland, & Goldberg, 1982), this suggests that they might struggle more with variability in tasks where adults seem to cope well, such as L2 vocabulary learning. Sinkevičiūtė et al. (2019) interpret their findings in terms of an account where there is a balance between the difficulties of processing input with varying talkers, and the potential benefits for generalisation. If variation places too much burden on the learner given the nature of the task and the learner's age, the benefit for generalisation may not be seen.

The key take away from the studies discussed here is that, contrary to what has been found for both adult L2 learners and some results with infants in the L1, high variability does *not* seem to be beneficial for children learning L2 vocabulary. This is similar to what has been found in the phonetic training literature discussed in

Section 1.3.5: the research directly comparing HV and LV input has found fairly consistent evidence of an HV benefit in adults, albeit in only a handful of studies. This benefit has not been clearly established in children, with one study finding evidence for a reverse LV benefit in generalisation in a discrimination task, while the other found evidence for an HV benefit in the same task (at least with novel items), but an LV benefit for generalisation to production. Thus, the evidence as to the role of input variability in child second language learners is not conclusive, warranting further research, which is the goal of this thesis.

1.5 Methodological and statistical concerns

In the last decade, there has been considerable focus in psychology on how methodological and statistical concerns might have led to published findings not being robust, known as the replication crisis. This introduction will finish by considering these issues with respect to the literature reviewed above.

One key concern of the replication crisis is publication bias, whereby null results are not published. There is some evidence that researchers might be less inclined to write up their null-results in the first place (Franco, Malhotra, & Simonovits, 2014). However, even when null results do get written up, they might still not make it into journals. This publication bias, where many journals are reluctant to publish non-significant results, is known as the file drawer problem. This problem was already identified forty years ago (Rosenthal, 1979) but unfortunately is still

prevalent today (Ferguson & Heene, 2012). This publication bias has been demonstrated in Masicampo & Lalande (2012), who find that in a set of psychology journals, there is an unusually large number of studies that have p-values just below .05, the traditional cut-off point of what is seen as significant results in null hypothesis significance testing. This could indicate a systemic problem of journals not publishing non-significant results, as a result of which researchers might feel driven to use questionable research practices, such as continuously adding to the sample until a significant p-value is obtained (data peeking – Francis (2012)); see John, Loewenstein, & Prelec (2012) for an estimate of the prevalence of questionable research practices. In terms of the literature reviewed above, this could be relevant for the finding of a variability benefit on generalisation, i.e. though most published studies seem to find this benefit, it is impossible to know how many other studies were conducted but did not find the effect.

Another concern is the problem of studies in psychology generally being under-powered (Maxwell, 2004). Dong et al. (2019) already raised the point that in phonetic training studies, a very large sample would be required to find evidence for an interaction between training condition and individual difference characteristics. In fact, it is likely that even powering the difference between HV versus LV input requires larger samples than are usually used in these training studies – studies that have directly compared input variability have had 15 participants per condition (Sadakata & McQueen, 2013, 2014), to around 20 per condition (Evans & Martín-Alvarez, 2016; Giannakopoulou et al., 2017; J. W. S.

Wong, 2012), through to maximally 30 participants per condition (Clopper & Pisoni, 2004; Perrachione et al., 2011). Therefore, the combination of possible publication bias and relatively small sample sizes may lead to inaccurate literature.

An additional problem in the phonetics training literature, and in psychology more broadly, is that where there are studies that do present a null result, the type of statistics used for inference (p-values) do not allow differentiation between ambiguous evidence where the data is insensitive to the hypothesis, and evidence against the hypothesis. This could mean that those studies might not have had enough power to show the effect they were looking for, as small effects require larger samples. Yet a common mistake in the literature is to interpret null effects as providing evidence against the hypothesis in question. With respect to the current literature, there have been examples of null results in various contexts, including studies that did not see a benefit of including orthographic symbols in training (Simon et al., 2010, Section 1.3.1), studies that did not find transfer from perception to production (Iverson et al. (2012); Shinohara & Iverson (2018, Section 1.3.3), and studies that did not see a benefit for younger learners (Giannakopoulou et al. (2017); Heeren & Schouten (2010); Wang & Kuhl (2003), Section 1.3.4). Unfortunately, it is not possible to draw conclusions about the original hypotheses from these null results. Only two of the studies discussed above used statistical methods which allowed the researchers to draw conclusions about the null: the studies by Dong et al. (2019) and Sinkevičiūtė et al. (2019). Specifically, these studies followed up non-significant results by using Bayes Factors, which can

distinguish between whether there is evidence for the null hypothesis (i.e. against the predicted hypothesis), or whether the evidence is ambiguous and the data are insensitive (Dienes, 2014).

The thesis addresses these concerns as follows. Firstly, while it aimed to use a larger sample than is generally used in the literature, working within the practicalities possible when testing children in schools meant that all experiments in Study 1 used 24 participants per condition (48 children and 48 adults per experiment), while in Study 2 the older age group was similar in size and the younger age group was slightly smaller. Although this is not necessarily a larger sample than is used in the literature, it is solidly on the upper end of the usual sample size. Secondly, this thesis uses Bayes Factors as the key inference statistic, allowing for quantification of the evidence for the null as well as for H1. Note that this also shows where there is insufficient evidence to evaluate the hypothesis (i.e. where the hypothesis is underpowered). In principle, it would be licensed to collect more data there (see Dienes (2016) for an explanation of how Bayes Factors eliminate the problem of optional stopping), but this was beyond the practicalities of the current thesis. Finally, for Study 2, this thesis uses pre-registration. Pre-registering studies in advance (Nosek & Lindsay (2018); Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit (2012)) has been suggested as one possible measure to combat publication bias (see Cumming (2014) for a more detailed discussion of other ways research should change). In doing so, researchers need to specify what analyses they intend to perform and what hypotheses they might have. Additional analyses

not stated in the preregistration are possible but will be considered exploratory, since the researchers had no initial hypothesis, and can thus not draw any strong conclusions. Study 2 was pre-registered at https://osf.io/kmspq.

1.6 Overview of the current thesis

All in all, second language learning in children is an important but under-researched area. This is likely because of the widespread assumption that children find learning a second language easier than adults based on the critical period hypothesis, but we have seen that there is no clear evidence for this with respect to phonetic training. Children might actually find it harder to learn than adults: in Shinohara (2014) the youngest children did not show any evidence of learning, while in most other studies younger children are worse overall. It is, therefore, important to consider which materials used in training children might elicit the most effective learning of L2 speech sounds. One important potential factor is high versus low variability training input, as HV input should aid generalisation, while at the same time children might find it harder. We have seen that the findings in the literature are currently conflicting as to whether to expect an HV advantage (Evans & Martín-Alvarez, 2016), an LV advantage (Giannakopoulou et al., 2017) or no variability effect (Sinkevičiūtė et al., 2019). However, there is more to consider. One question is whether orthography, or orthographic type symbols, are helpful for children. We have seen that most training studies with adults have made use of these symbols, and there is some evidence that they are useful for these learners

(Escudero et al., 2008), however, we have seen that using orthography may cause difficulty for child learners (Shinohara, 2014). This thesis will explore the effect of high and low variability input in phonetic training for non-native speech sound learning in children, aiming to find a child-friendly adaptation of the existing HVPT paradigm. As a secondary goal, it will investigate the role of using orthography or orthography-like symbols in the training materials. This thesis will investigate these factors by looking at English children and adults learning Dutch vowels, and Dutch children learning English vowels. In the analyses, Bayes Factors as well as frequentist analyses are used: Bayes Factors allow for evaluation of the evidence for the null, and therein overcome the problems of overinterpreting the absence of an effect based on traditional frequentist statistics.

Chapter 2 provides a detailed description of the phonetics of English and Dutch, as these are the languages that are used in the empirical chapters that follow. The chapter starts by detailing the general phonetic inventory of each language to provide the phonetic context, before specifically discussing the vowel contrasts that are known to be hard for Dutch learners of English and English learners of Dutch, respectively.

Chapter 3 is the first of two experimental chapters, and explores the use of the HVPT paradigm in L2 speech sound learning in children by training native English speaking adults and 7-8 year-old children on Dutch vowel contrasts. Here, adults and children are both naïve learners of the target language and are thus matched in their L2 exposure, unlike many previous studies where adults have been learning

the language for a longer period than the children. The first experiment aimed to investigate the role of high or low input variability using a paradigm in which participants were trained using an identification task involving orthography-like symbols (since these novice participants had no knowledge of actual Dutch orthography). Since learning was low, particularly in children, the focus of the study shifted to developing an HVPT task that would lead to substantial learning and generalisation in children, with the hope that this would make the detection of differences between the variability conditions more likely. Although the predicted HV benefit was not found in any of the three experiments, with substantial evidence for the null in some tests with both learner groups, an interesting result was found, in that learners showed better learning with picture-based stimuli than with the orthography-like symbols.

Chapter 4 presents a large-scale experiment which aims to follow up some of the questions arising in Chapter 3: since learning in Study 1 was quite low, a longer training study was employed, as is more commonly used in the literature. In order to feasibly be able to implement a longer training paradigm, this required working with participants who were already learning the language, in a context where language learning in school is highly valued as they needed to be willing to give up a substantial amount of time to take part in the study. Therefore, Study 2 used native Dutch 7-8 year-olds and 11-12 year-olds, who were trained on English vowel contrasts over two weeks. Their learning and generalisation abilities were tested through means of a battery of pre/post-test perception tasks, vocabulary

tasks, and a production task. Orthography and picture identification were tested separately to pull apart the individual effects. All tasks investigated generalisation to novel voices, while some also tested generalisation to novel items. Additionally, performance of the two age groups was compared to explore potential age differences. Results showed that although both groups improved in training, only older children managed to generalise to novel voices at test, with evidence for the null in younger children, at least in some tasks. Critically, there was no evidence for a benefit of HV input on generalisation.

Finally, Chapter 5 summarises the main findings of the studies conducted in this thesis. Key topics include the role of orthography versus pictures in training, the different effects seen in different test tasks, age differences seen across the studies, and finally the key finding of the thesis: the consistent absence of a benefit of high variability in generalisation, for any age group in any of the tasks, with substantial evidence for the null in many cases. The thesis finishes by considering the implications of these findings and directions for future research.

Chapter 2. The phonetics of English and Dutch

This chapter will give a short overview of the basic phonetic properties of both English and Dutch for the respective varieties used in the rest of this thesis. In addition to giving an overall phoneme inventory, it will focus on the particular vowel contrasts of interest to this research, namely those contrasts that are hard for English learners of Dutch, and Dutch learners of English. The two languages will be discussed in order of relevance to the rest of the thesis, with the first section describing the sounds of Dutch for English learners, relevant for the experiments of Study 1 in which English speaking adults and children were taught Dutch vowels, followed by the section describing the sounds of English for Dutch learners, relevant for the training study in Study 2 where Dutch primary school children were learning English.

2.1 The sounds of Dutch for English learners

There have been several descriptions of the phonetics and phonology of Dutch in varying amounts of detail (see e.g. Booij (1999); Gussenhoven (2002)), although very few of these are aimed at English learners (Collins & Mees, 2003). However, all of these descriptions tend to describe a form of 'Standard Dutch' based on the variety mostly heard in broadcasting (*Algemeen Nederlands* or the more archaic *Algemeen Beschaafd Nederlands* in Dutch). While this is said to be a fairly common variety heard in everyday life such as in schools or governmental bodies, this is

only the case for some parts of the Netherlands. When going outside of the *Randstad* (the major cities in the north-west of the country) speakers will most often use a more regional variety of Dutch, or a hybrid variety with regional variants alongside standard variants. This is also the case for the variety of Dutch used in the research for this thesis. The particular variety of Dutch described here is that used by both the speakers used to record the Dutch stimuli in Study 1, as well as by the Dutch primary school children participating in the training study of Study 2. This variety is a southern variety of Dutch known as East Brabantian, spoken in the eastern part of the Noord-Brabant province. In particular, all were speakers of the *Kempenlands* sub-variety as defined by the geographical division of accents and dialects in the *Woordenboek van de Brabantse Dialecten* ('Dictionary of Brabantian Dialects', Belemans & Goossens (2000)).

It is important to note that despite the fact that all Dutch speakers and participants came from the same area, there will likely still have been some structural variation between them. In particular, the most likely source of variation in this case will be between the Dutch speakers who recorded the stimuli for Study 1, and the participants for Study 2: most of the child participants in Study 2 will have spoken a more urban variety of *Kempenlands* as the school they were recruited from was in a major city, and none of the children indicated speaking the local dialect (as is more often the case in urban centres and younger generations in recent years; Swanenberg (2009) already describes the downfall of the dialect in his inaugural lecture). On the other hand, the speakers recorded for Study 1 all grew up in more

rural areas to the west and south-west of this city where the local dialect remains more common. However, there are no sources that describe systematic distinctions between the more rural and more urban varieties; in fact, even the rural area has a substantial amount of variation that enables speakers to be identified as being from a particular village depending on their pronunciation as well as what vocabulary they use. Since most of this variation is not related to the vowel system, which is the main interest of this thesis and is the focus of the stimuli recorded for Study 1, it did not seem problematic to use speakers who grew up in different villages within this area. It is important to keep in mind that while most of the speakers recruited for stimuli recording of Study 1 do in fact also speak their local dialect, the speakers were recorded speaking standard Dutch lexis with a regional accent, and not the local *Kempenlandse* dialect. The latter has many vowels that are quite distinct from Standard Dutch, and distinctive feature combinations can differ significantly from village to village (for a more detailed introduction to some of this variation in the *Kempenlandse* dialect, see de Bont (1962); Swanenberg & Swanenberg (2002)).

Note that there is some controversy surrounding the transcription of some vowels of Dutch, where descriptions of Dutch diverge on several points (see e.g. Booij (1999); Collins & Mees (2003); Gussenhoven (2002, 2007); Gussenhoven & Broeders (1997)). Two disputes about the transcriptions of vowels are briefly expanded on here, and the choice of transcription used in this thesis is justified. The first debate is on the transcription of the initial element in the back diphthong as either /ɔu/ or /ɑu/. Historically, the initial element was rounded, and it still tends

to be in Belgian varieties of Dutch where /ɔu/ is used, but in Netherlandic varieties this rounding is much less present. In this thesis /ɑu/ will be used as the initial element is clearly unrounded in the diphthong used in the East Brabantian accent, so it is better suited to the actual realisation. The second dispute concerns the transcription of the short close rounded vowel (either /ʉ/, /ɵ/ or /ʏ/); historically, this vowel has been transcribed as /œ/, but more recently other transcriptions have come into use, all argued to represent slightly different versions of the vowel. In this thesis /ʏ/ will be used, based on the argument posed in Gussenhoven (2007) that the vowel is more like a rounded equivalent of /ɪ/ than of /e/ or /ɛ/.

2.1.1 Consonants

See Table 1 for an overview of the consonant inventory of East Brabantian Dutch. East Brabantian, like other varieties of Dutch, does not distinguish between voiced and voiceless obstruents in syllable-final position (i.e. it has final devoicing). There is currently a sound change in progress in which voiced fricatives lose their voicing in onset as well, meaning not all speakers make a distinction between /s/-/z/ or /f/-/v/ in word-onset. As this change is spreading across the country from what seems to be the *Randstad* outwards, it is as of yet not stable enough to describe how advanced it is in Noord-Brabant. East Brabantian Dutch also shows a rich variation in allophonic variation of /r/, as is well-known to be the case in Dutch (see van de

Velde & van Hout (1999, 2001); see also Sebregts (2015) for a recent sociophonetic investigation of Dutch /r/ in which 20 different variants of /r/ are distinguished).

Although East Brabant is a relatively small and well-defined area, there is still quite a lot of /r/-variation; key factors that seem to play a part in which allophones speakers use are age, register, and geographic location. Verstraeten & van de Velde (2001) find that in the Netherlands, individuals tend to use up to three different variants of /r/. The uvular trill, approximant, and fricative are generally considered the most common in the area of interest, though note that an alveolar tap or trill is also attested, particularly in older speakers (Swanenberg & Swanenberg, 2002).

A salient feature typically associated with the more general Noord-Brabant accent is the use of the velar fricatives /x/-/ɣ/ for <g>, commonly described as having a *zachte g* ('soft g'). Standard Dutch and most varieties roughly north of the three main rivers in the Netherlands are described as having a *harde g* ('hard g'), and generally use the voiceless uvular fricative /χ/ in all positions. The use of the *zachte g* over the *harde g* is a very salient and marked feature, and listeners are very aware of its geographical and social connotations. A less salient feature that also identifies speakers as being from the south of the Netherlands is the use of the voiced bilabial approximant /β̞/ rather than the voiced labiodental approximant /ʋ/ for <w> (Booij, 1999, p. 9; Collins & Mees, 2003, p. 175; Gussenhoven & Broeders, 1997, p. 73). Furthermore, East Brabantian Dutch also has consistent schwa-epenthesis in non-

homorganic consonant clusters in coda where the first element is either /r/ or /l/, and the second element is not /s/ or /t/, such as in *markt* 'market' /ˈmɑrəkt/ or *melk* 'milk' /ˈmelək/. It seems to be one of the varieties of Dutch where the epenthetic schwa is consistently produced, potentially due to influence of the dialect, which also has this feature (see Kloots, De Schutter, Gillis, & Swerts (2002); Kloots, Gillis, De Maeyer, & Verhoeven (2012) for a more detailed discussion of schwa-epenthesis in varieties of Dutch). In contrast, in other varieties of Dutch schwa-epenthesis is described to be optional (Booij, 1999; Kirstein, 2018; Warner, Jongman, Cutler, & Mücke, 2001). Another common process in the East Brabantian variety of Dutch is /t/-deletion at the end of frequent mostly function words where other varieties of Dutch would pronounce the /t/, as in for example *wat* 'what' /βɑ/ and *niet* 'not' /ni/.

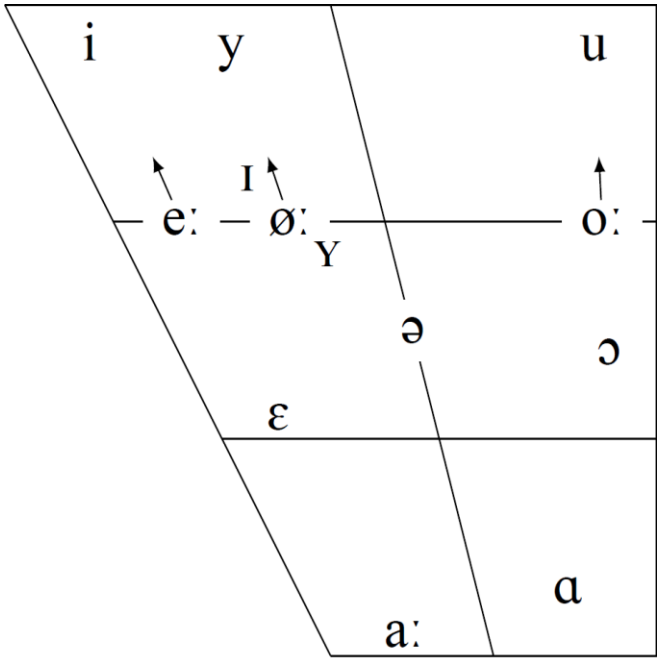
	Bilabial		Labiodental		Alveolar		Post-alveolar		Palatal		Velar		Uvular		Glottal	
Plosive	p	b			t	d					k	(g)			(ʔ)	
Nasal		m				n				(ɲ)		ŋ				
Trill						r[1]							ʀ̥[1]	ʀ[1]		
Tap or Flap						(ɾ)[1]										
Fricative			f	v	s	z	(ʃ)	(ʒ)	(ç)	(ʝ)	x	ɣ	(χ)	ʁ[1]		ɦ
Affricate																
Approximant	β̞									j				ʁ̞[1]		
Lateral approximant						l										

Table 1. Consonant inventory of East Brabantian Dutch, with the voiceless phoneme on the left and the voiced phoneme on the right of each column. Phonemes in brackets are marginal and tend to only occur as allophones. All phonemes marked [1] are variants of /r/ possibly used in the area; note that speakers will not necessarily use all of these variants, but depending on phonetic context as well as where they are from, they may use a subset of these. Based on Booij (1999); Collins & Mees (2003); Gussenhoven & Broeders (1997); van de Velde & van Hout (1999).

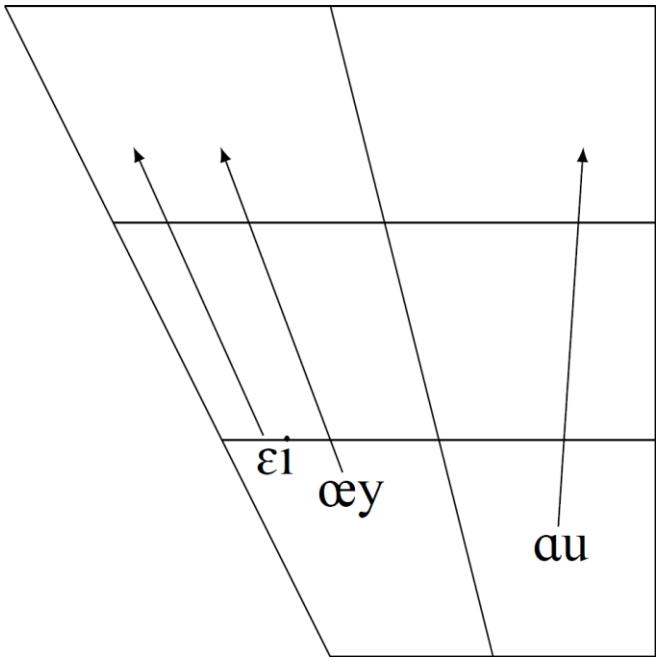
2.1.2 Vowels

The East Brabantian variety of Dutch has eight short monophthongs, schwa, and four long monophthongs three of which are undergoing diphthongisation (/eː/, /øː/, and /oː/, as indicated by the arrows in Vowel diagram 1 below. Note that /øː/ can move towards /w/ or towards /j/, depending on the regional variety and speaker; in the area the speakers were recorded it goes towards /j/ making it more like a fronting diphthong). There are also three true diphthongs /ɛi/, /ɑu/, /œy/ (see Vowel diagram 2). Additionally, there are up to eleven marginal vowels that are mainly used in French loanwords; due to their fairly low frequency, these will not be discussed here.

In addition to these diphthongs, there is an additional set of what are sometimes called semi-diphthongs or vowel sequences, /aːi/, /oːi/, /ui/, /iu/, /yu/, /eːu/, as well as two rare diphthongs, /ɔi/ and /ɑi/. There is no agreement on whether these should be analysed as diphthongs or vowel sequences, so for the purpose of this thesis they will not be discussed in more detail.



Vowel diagram 1. Monophthongs of East Brabantian Dutch. Based on Booij (1999); Gussenhoven (2002, 2007); Gussenhoven & Broeders (1997).



Vowel diagram 2. Diphthongs of East Brabantian Dutch. Based on Booij (1999); Gussenhoven (2002, 2007); Gussenhoven & Broeders (1997).

It is important to keep in mind that all of these vowels are coloured by the way they are spoken in the local dialect, in particular since most of our recorded speakers also spoke the dialect themselves. For this reason, this section will now go into more detail on the vowel differences in the East Brabantian dialect compared to a more standard variety of Dutch, and will point out how this may have influenced the production of Dutch with an East Brabantian accent. Main trends will be described, but note that not all phenomena are easy to capture in a few sentences; where possible the specific context in which they happen has been provided if this was deemed feasible in the amount of space available, but priority was given to describing the resulting realisation of the vowels over the specific context. For a more detailed discussion of all of these phenomena and the overall dialect, see for instance Belemans & Goossens (2000); de Bont (1962); Goossens (1981); Swanenberg & Swanenberg (2002), and see Swanenberg & Brok (2008) for a more detailed bibliography. Note that the English listeners who were trained on these vowels had not been exposed to any Dutch before the experiments in this thesis; this means that to them, the regional differences will not be relevant as they do not have any knowledge of what different regional pronunciations or a non-regional standard might sound like.

The East Brabantian dialect typically shortens phonologically long vowels, of which standard Dutch has /aː/, /eː/, /øː/, and /oː/. Of these, /aː/ is often realised as [ɒː] or [ɔː] in East Brabantian, which can be shortened to something akin to /ɒ/ or

/ɔ/. /eː/, which tends to be realised as [ɛː], becomes /ɛ/ or even /ɪ/, while /øː/ tends to become [ʏ] when shortened. The long back vowel /oː/, which can be realised as [ɔː], seems to be less affected by this process; potentially this is the case to retain the contrast with the shortened realisation of /aː/. At the same time, the East Brabantian dialect also lengthens phonologically short vowels. The process is mostly context-dependent: it occurs in closed syllables before /x/ or /s/ and /st/. The lengthening can also lead to some diphthongisation in the case of lengthened /i/, /u/, and /y/ (de Bont, 1962; Swanenberg & Swanenberg, 2002). The combined phenomena of shortening long vowels and lengthening short vowels means that the vowel length for long and short vowels might differ in an East Brabantian accent, compared to standard Dutch.

Another process that happens in the East Brabantian dialect is umlaut: the fronting and/or raising of vowels. The umlaut that happens in East Brabantian is partially structurally caused by the grammar, where it is for instance consistently applied in creating plurals, the third person singular of verbs, and diminutive nouns. For example, the first person singular for 'walk' in the present tense is *loop* /lɔːp/, but in the third person this becomes *lupt* /lʏpt/. The part of the umlaut process that is not explained by grammatical rules is due to the historical sound change that occurred in Germanic languages, sometimes known as *i-umlaut*: this affects vowels when they are (or were historically) followed by a syllable containing /i(ː)/. This

process is generally described in two stages: the first stage happened around the 500s in all of the West-Germanic languages, including Dutch, but the second only happened starting from the turn of the 11th century where it no longer applied to all West-Germanic languages. This second i-umlaut stage is typically understood to not have occurred in standard Dutch, but it did occur in the East Brabantian and Limburgian dialects spoken in the south east of the Netherlands (Goossens (1981), see de Bont (1962, p. XLIII) for a more detailed description of when the umlaut occurs). This *i-umlaut* in East Brabantian leads to /oː/ becoming /øː/ (which then can get shortened to [ʏ] as per the shortening process described above), while /ɑ(ː)/ becomes [æ(ː)] or [ɛ(ː)]. /ɔ/ can become [ʏ], and /u/ becomes [y(ː)]. One example of this is the word *groen* 'green', which in standard Dutch is pronounced as /χrun/ with a back vowel, but in East Brabantian becomes /ɣʁyn/ with a front vowel instead. The effect umlaut might have on East Brabantian speakers of Dutch is that they might use slightly more fronted or raised variants of the vowels that would have been affected by this process in the dialect.

Finally, the fronting diphthongs tend to be monophthongised in East Brabantian (de Bont, 1962): /œy/ can be realised as [œː], and /ɛi/ tends to be realised as [ɛː]. Since these vowels can then be analysed as long vowels rather than diphthongs, this means they are also susceptible to the shortening of long vowels, resulting in realisations of [œ] and [ɛ] respectively. Speakers might thus use a more

monophthongised version of the /œy/ and /ɛi/ diphthongs in the East Brabantian accent when compared to standard Dutch. In general, the pronunciation of /œy/ is said to show a lot of variation even within the East Brabantian area, and within the small area of East Brabant, it can be realised as [œy], [œː] or [œ], [øː], [ɑu], [oː], [ɑː], or [yː] (Swanenberg & Swanenberg, 2002). The Kempenlandse variety has the [œː] realisation in its dialect, which might indicate that the vowel /œy/ might be realised as having a longer initial element of the diphthong by our speakers compared to a more standard variety of Dutch.

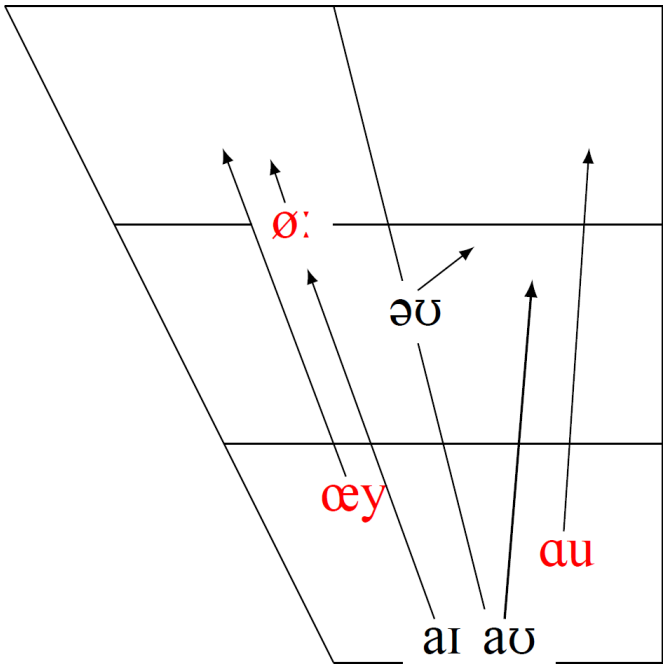
Confusing vowel contrasts

Since there is very little literature on second language acquisition of Dutch by English learners, a pilot study was conducted which aimed to determine the contrasts that might prove difficult for these learners. In this pilot, 10 native speakers of Standard Southern British English with no experience of Dutch performed an auditory three-way discrimination task where they were asked to select the odd one out. Listeners heard naturally recorded East Brabantian accented Dutch vowels in a /h/-V-/t/ context. Fifteen different words were recorded, some of which were real Dutch words and some of which were nonsense words (though note that for the native speakers, who were naïve learners of Dutch, this distinction would not matter). All Dutch monophthongs and diphthongs (apart from schwa) were included: *hit* /ɦɪt/, *hiet* /ɦit/, *huut* /ɦyt/, *hut* /ɦʏt/, *haat* /ɦaːt/, *had* /ɦɑt/, *heet*

/ɦeːt/, *het* /ɦɛt/, *heut* /ɦøːt/, *hoed* /ɦut/, *hoot* /ɦoːt/, *hot* /ɦɔt/, *hout* /ɦɑut/, *huid* /ɦœyt/. All possible minimal pairs of the 15 vowels were included in the task and each minimal pair combination was encountered once, resulting in 105 trials in total.

On the whole, error rates were low; all participants made at least two errors overall, but none made more than nine errors. Nine different vowel contrasts resulted in errors, all for at least two or more people. Of these contrasts, the two with the highest error count were <heut-huid> /ɦøːt/-/ɦœyt/ with five errors, and <hout-huid> /ɦɑut/-/ɦœyt/ with four errors. Since the contrast hout-heut /ɦɑut/-/ɦøːt/ was also relatively high in terms of error (two errors), the three-way contrast /øː/-/ɑu/-/œy/ was chosen for use in the training study reported in the experiments in Study 1 (see Vowel diagram 3). Note that although a three-way contrast might be cognitively more demanding to acquire than a two-way contrast, particularly for child participants, there was a good reason to make this choice: training learners on all three contrasts provides them with a clearer overview of the vowel space than a two-way contrast would. Since native English speakers are especially prone to confuse both /øː/ and /ɑu/ with /œy/, providing them with anchoring points for the former two diphthongs (which they can tell apart from each other slightly more easily) should be helpful in also getting them to adjust their representation of /œy/, and might be especially helpful for children as it provides them with more

redundant cues, and thus certainty, as to the identity of the diphthongs they are being trained on.



Vowel diagram 3. Confusing vowels for English learners of Dutch. East Brabantian Dutch vowels indicated in red, Standard Southern British English closest equivalents or common substitutions indicated in black. East Brabantian Dutch vowels based on Booij (1999); Gussenhoven (2002, 2007); Gussenhoven & Broeders (1997); SSBE vowels based on Carley et al. (2018); Cruttenden (2008); Roach (1991); Wells (1982b).

It is perhaps not surprising that /œy/ shows up among the contrast with the highest error rates, as this vowel is comprised of two elements that are not part of the Standard Southern British English phoneme inventory. The diphthong itself has anecdotally been called 'characteristically Dutch', and has been said to be linguistically unusual as it is fairly rare in other languages (Mees & Collins, 1983; Warner, 1999). The difficulty with /øː/ is also not unexpected, as this vowel has no close Standard Southern British English equivalent. The confusion between these

two vowels is, therefore, easily explained: both vowels show similar fronting towards the second element, in particular due to the diphthongisation in progress for /øː/, and the key difference between the vowels is found in the starting point of each of them with that for /œy/ being a more open-mid vowel with a large amount of fronting while /øː/ starts close-mid and has less fronting.

More surprising is the confusion between /ɑu/ and /œy/. In principle, these diphthongs are quite distinct: /œy/ is a fronting diphthong, while /ɑu/ is a backing diphthong, and their starting points are indeed also in line with that. However, one possibility is that a phenomenon known as *GOOSE-fronting* might play a role. *GOOSE-fronting* is a vowel change in progress in British English that involves the traditionally back close rounded vowel /u/ fronting, for which the amount of fronting is currently still speaker-dependent (Cruttenden, 2008). However, for a lot of speakers, the vowel is now produced somewhere between central and fully front, and can in fact be quite close to the front close rounded vowel /y/. This means that there is quite a lot of variation in the production of /u/ an English speaker may be exposed to. Since pronunciation variants such as [u] and [y] assimilate to their native representation of /u/, this could result in English learners of Dutch hearing the second element of /œy/ and interpreting it as a back rounded vowel /u/, making the diphthong sound much more like a backing diphthong

rather than the fronting diphthong it is. Since the initial element is midway between the central /ə/ and the open /a/, it could be understood as a variant of either SSBE /aʊ/ or /əʊ/.

Another possible explanation for the confusion between /ɑu/ and /œy/ is existing regional variation of the pronunciation of the English MOUTH vowel. For SSBE, this diphthong is pronounced [aʊ], with a fairly front open first element moving towards a back mostly close second element. However, there are other regional accents of English that produce a variant of MOUTH that is much closer to the Dutch vowel quality of /œy/. Most notably, in Northern Irish accents, the second element of the MOUTH diphthong is fronted and can lose its rounding. In Derry~Londonderry this diphthong is produced as [ɔ̈y] (McCafferty, 1999), which is extremely close to the vowel quality of the Dutch vowel. A similarly fronted second segment is found in Belfast, where the MOUTH diphthong is generally said to be produced as [əʉ] (though note the production of the MOUTH vowel is extremely variable in Belfast because of its use as a sociolinguistic marker, Wells (1982b)). Some variants of Scottish English, particularly Glasgow (Stuart-Smith, 1999) and Edinburgh (Chirrey, 1999) are also described as having a close fronted second element, producing MOUTH as [ʌʉ], though the fronting is less extreme than in the Northern Irish accents. Taking this regional variation of the MOUTH vowel into account could explain why the Standard Southern British English

speakers conflated the Dutch /ɑu/ and /œy/ vowels. While they have not previously encountered the Dutch vowels, they may have previously been exposed to regional variants of English that produce the MOUTH diphthong with a fronted second element, despite it being a backing diphthong in Standard Southern British English. Even if speakers have not had much exposure to these varieties of English, they could still be aware, whether implicitly or explicitly, that a fronting diphthong is a regional variant of MOUTH. If this is the case, they might be inclined to map both the fronting /œy/ as well as the backing /ɑu/ onto their representation of the English MOUTH vowel, leading to difficulties in telling the two diphthongs apart.

Looking at the relative difficulty of the various vowel contrasts in light of PAM, the contrast between /œy/ and /øː/ will likely result in single category assimilation to a native English diphthong (the most probable candidates being /əʊ/ or /aɪ/), possibly with a category goodness difference. Similarly, /œy/ and /ɑu/ will likely result in a single category assimilation with a category goodness difference (most probably assimilating to SSBE /aʊ/). PAM would predict the latter contrast to be easier to distinguish than the former as the Dutch articulatory gestures are more distinct, but the behavioural results from the pilot and further explanation discussed above suggest that for SSBE listeners, both seem to result in similar amounts of difficulty. On the other hand, the contrast between /ɑu/ and /øː/ will

more likely result in two category assimilation (probably mapping the former onto /aʊ/ and the latter onto /əʊ/ in the first instance); this should make this contrast easier to perceive and acquire than the other two.

2.2 The sounds of English for Dutch learners

Dutch people are often said to speak good English, and articles about this frequently make the rounds on popular media. In a survey by the European Commission, 90% of Dutch people interviewed said they were able to speak English to at least a conversational level (TNS Opinion & Social, 2012). Part of this seeming ease in speaking English is to do with the fact that English and Dutch are both Germanic languages and have a considerable amount of overlap in grammar, vocabulary, and pronunciation. Additionally, English is an obligatory subject in Dutch primary and secondary schools, and in general English lessons start from the age of 10, although there is currently a pilot ongoing where they start from the age of four (EP Nuffic, n.d.). These lessons mainly focus on being able to communicate in English rather than on for instance rote learning grammatical rules. Moreover, English is also increasingly used in the media, exposing Dutch people to many different varieties of English, with English-spoken television generally being subtitled rather than dubbed unless it is aimed at young children. All of these factors contribute to the relatively high English fluency of the Dutch population.

The target variety of English that will be used in this thesis is British English, because although Dutch people are exposed to many different varieties of English nowadays, in education British English is generally the teaching model of choice (van der Haagen, 1998). While Received Pronunciation (RP) may still be used as a model in higher education teaching in the Netherlands, often the reality of the variety learners aim for is in fact not traditional RP as it is described in for instance (Wells, 1982b, 1982a)). With even Queen Elizabeth having changed her pronunciation over several decades of Christmas broadcasts (Harrington, 2006; Harrington, Palethorpe, & Watson, 2000), aiming for the traditional RP as it was spoken many years ago has strong social connotations and can make learners' accents sound affected. Rather, the variety many higher education institutions aim for is a modernised version of what was traditionally called RP, which includes many features of what is described in Wells (1982b) as 'Near-RP'. This modern variety has become much more common in recent years, also in the media, and is therefore considered a more representative model for what Dutch learners of English will be aiming for (see Cruttenden (2008) for a more in-depth discussion on contemporary RP and its variants, as well as changes in progress). Note that there is an ongoing debate on the terminology used to describe this more modern variety (see for instance Carley, Mees, & Collins (2018, p. 6) for an overview of terms used). This thesis will use the term Standard Southern British English (SSBE) for the variety of British English described.

2.2.1 Consonants

See Table 2 for an overview of the consonant inventory of Standard Southern British English. Generally, the consonants are less of a problem of Dutch learners of English, as there is a lot of overlap between the two consonant inventories. The exception is the lack of a Dutch equivalent for the SSBE dental fricatives, which leads to some issues in acquiring the correct pronunciation for these phonemes. Common substitutions are alveolar fricatives /s/ and /z/ or stops /t/ and /d/ for beginners (Collins & Mees, 2003), while more advanced learners might substitute labiodental fricatives /f/ and /v/ instead (Gussenhoven & Broeders, 1997). More prominent differences are to be found in the coarticulation rules and connected speech processes of Dutch and English. One particular feature that is most affected by this in terms of differences between English and Dutch is the voicing. Dutch is known to have lexicalised syllable-final obstruent devoicing (Collins & Mees, 2003, p. 53), while English does have a voicing contrast in syllable-final position (though note allophonic (partial) obstruent devoicing can occur phrase-finally as well as before or after a pause or voiceless phoneme, Carley et al. (2018, p. 14)). When applying the Dutch syllable-final obstruent devoicing rule to English, this can effectively neutralise a meaningful contrast, leading to misunderstandings between for example /bæt/ and /bæd/. In addition to this systematic phonological rule, Dutch also has voicing assimilation. In general, assimilation of voice does not occur as such in SSBE, but it is very common in Dutch, where regressive voicing,

progressive devoicing, and intervocalic voicing are all common results of coarticulation (Gussenhoven & Broeders, 1997). Sometimes, the difference between SSBE and Dutch is not just in one of the languages having an assimilation rule while the other does not; the assimilation processes can also go in the exact opposite directions in identical contexts. An example of this is regressive voicing of obstruents before /b/ or /d/ in Dutch, where SSBE would have progressive devoicing, such as Dutch *update* [ˈʏbdeːt] versus English 'update' [ˈʌp̬deɪt].

	Bilabial		Labiodental		Dental		Alveolar		Post-alveolar/ Palato-alveolar		Palatal		Velar		Glottal	
Plosive	p	b					t	d					k	g		
Nasal		m						n						ŋ		
Fricative			f	v	θ	ð	s	z	ʃ	ʒ					h	
Affricate									tʃ	dʒ						
Approximant										ɹ		j		w[1]		
Lateral approximant								l								

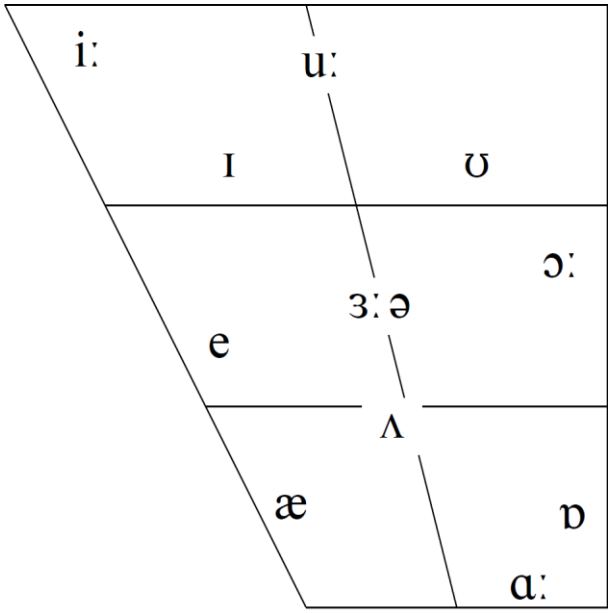
Table 2. Consonant inventory of Standard Southern British English, with the voiceless phoneme on the left and the voiced phoneme on the right of each column. [1] /w/ has a double articulation and is in fact a voiced labial-velar approximant. Based on Carley et al. (2018); Cruttenden (2008); Roach (1991).

2.2.2 Vowels

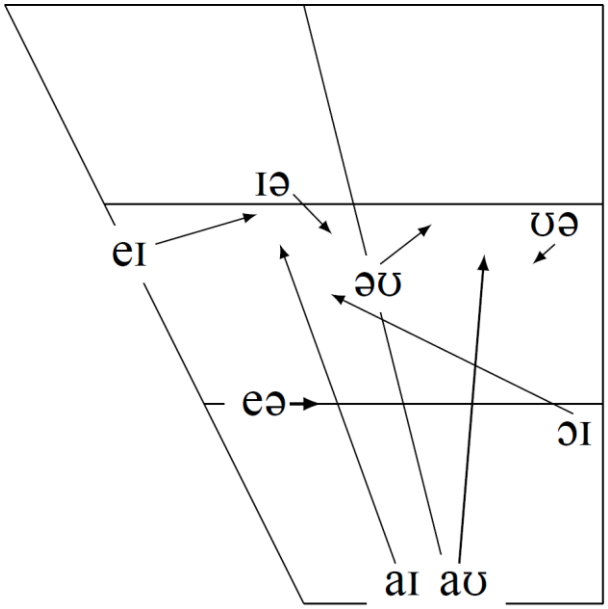
SSBE has six short monophthongs, schwa, five long monophthongs (see Vowel diagram 4), and eight diphthongs (see Vowel diagram 5). Three of those diphthongs, /ɪə/, /eə/, and /ʊə/, are undergoing monophthongisation (Cruttenden, 2008), shown in the diagram with a relatively shorter arrow. In addition to these diphthongs, there is an additional set of what are sometimes called triphthongs or diphthongal sequences, /eɪə/, /aɪə/, /ɔɪə/, /əʊə/, and /aʊə/. These vowels can undergo smoothing, where the second element is omitted. However, these vowels are not generally considered to be true triphthongs, but rather diphthongs followed by a separate schwa.

As previously mentioned, one change in progress to take note of is */uː/-fronting*. Originally, /uː/ is a close back rounded vowel, but over the past few decades it has been moving forward. There is currently still considerable variation in people's realisation of /uː/, which can be realised anywhere between its original back position (although this may sound marked) and a fully fronted vowel more similar to the close front rounded /y/. Younger speakers of SSBE are said to have a more fronted realisation than older speakers. Note that the SSBE speakers who were recorded (or who participated) in the studies in this thesis were mostly university students. Inspection of the recordings revealed that, as expected, the speakers

produced centralised variants of /uː/ close to [ʉː] (see Appendix I for the full formant measurements).



Vowel diagram 4. Standard Southern British English monophthongs. Note that /uː/ is depicted in a more fronted realisation. Based on Carley et al. (2018); Cruttenden (2008); Roach (1991); Wells (1982b).

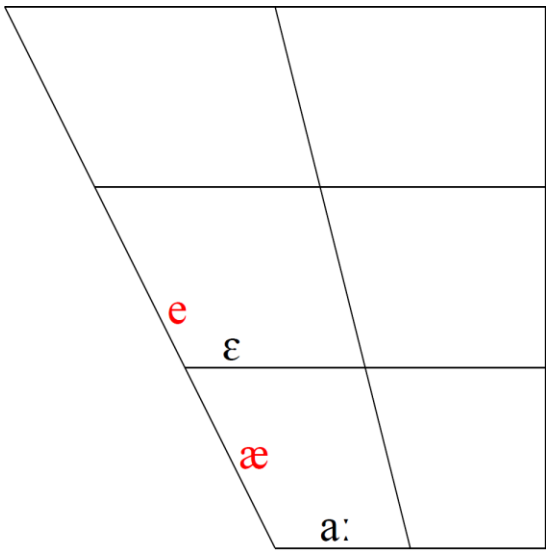


Vowel diagram 5. Standard Southern British English diphthongs. Note that the /ɪə/, /eə/, and /ʊə/ are all undergoing monophthongisation; this has been indicated by shorter arrows compared to the other diphthongs. Based on Carley et al. (2018); Cruttenden (2008); Roach (1991); Wells (1982b).

Confusing vowel contrasts

There is a large literature on English pronunciation for Dutch learners (see e.g. Collins, den Hollander, Mees, & Rodd, 2011; Collins & Mees, 2003, 2012; Gussenhoven & Broeders, 1997; Scott Sheldon, 1974; Willems, 1982). Most of this tends to be used in university teaching, and focusses on how learners can achieve the desired pronunciation for each English phoneme from a Dutch perspective. On the basis of this literature, the current section will focus on three contrasts that are used in Chapter 4 of this thesis, which are notoriously difficult contrasts for Dutch learners of English. This section will describe why these contrasts are particularly difficult in light of the learners' native language.

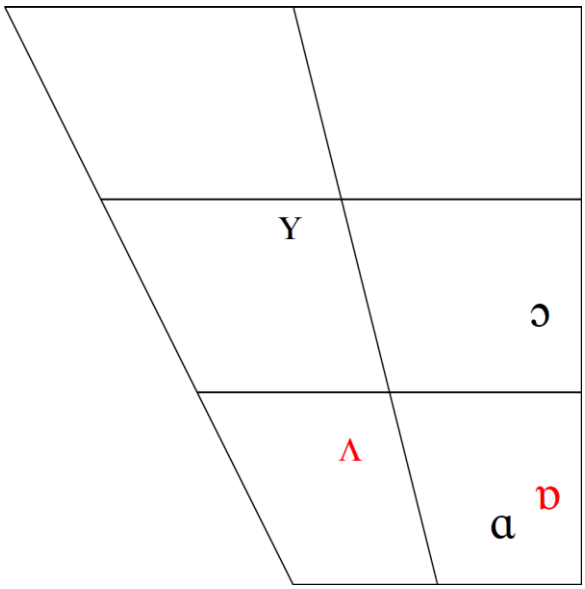
Note that the SSBE /iː/-/ɔː/ contrast, which is used in Study 2 as a control contrast, is a very straightforward contrast to acquire for Dutch learners of English. Both vowels have a Dutch equivalent that is very close if not in roughly the same position in the mouth (with /i/ as the equivalent for /iː/, and /oː/ as the equivalent of /ɔː/, with the marginal vowel /ɔˑ/ from words such as *roze* 'pink' also being an option depending on consonant context). There are some minor differences in vowel quality, but these have more to do with differences due to coarticulation and connected speech processes, and generally do not lead to problems in perception or production.



Vowel diagram 6. Commonly confused SSBE /e/-/æ/ vowel contrast, with Dutch closest equivalent vowels. SSBE vowels indicated in red, East Brabantian Dutch closest equivalents or common substitutions indicated in black. East Brabantian Dutch vowels based on Booij (1999); Gussenhoven (2002, 2007); Gussenhoven & Broeders (1997); SSBE vowels based on Carley et al. (2018); Cruttenden (2008); Roach (1991); Wells (1982b).

The /e/-/æ/ contrast is the most well-known of the challenging contrasts for Dutch learners of English, and most learners are aware of this. As can be seen in Vowel diagram 6 above**,** /e/ is relatively close to the Dutch /ɛ/, and learners can (and do) use their Dutch vowel for the English target without it sounding too conspicuous or causing problems for comprehension. This may in part be due to the production of /e/ being variable in English, with some accents, such as General American, Welsh English, and Yorkshire English, having more open variants closer to [ɛ] (Wells, 1982b, 1982c). However, for /æ/, there is no close equivalent in Dutch. The vowel maps onto the vowel space between /ɛ/ and /aː/; the latter of these is a long open front vowel, where the length distinction does not make it a suitable alternative for SSBE /æ/. This means that beginners tend to collapse the /e/-/æ/

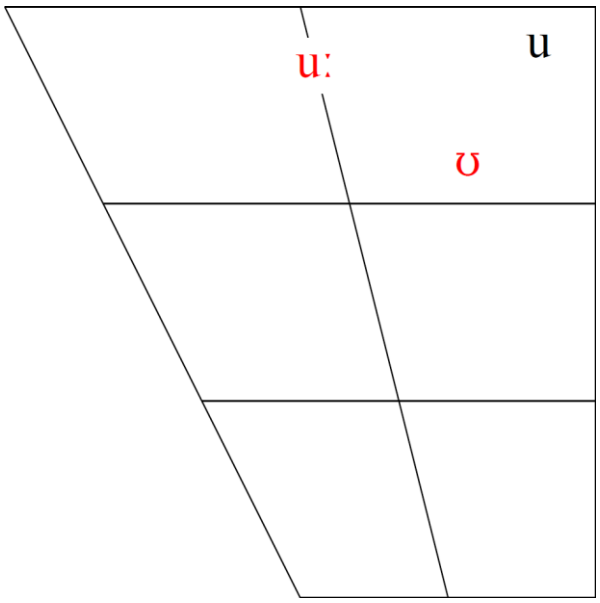
contrast onto their native /ɛ/, effectively making *bed* and *bad* indistinguishable from each other. Since this is a fairly common and productive minimal pair contrast in the English lexicon, this can lead to confusion.



Vowel diagram 7. Commonly confused SSBE /ʌ/-/ɒ/ vowel contrast, with Dutch closest equivalent vowels. SSBE vowels indicated in red, East Brabantian Dutch closest equivalents or common substitutions indicated in black. East Brabantian Dutch vowels based on Booij (1999); Gussenhoven (2002, 2007); Gussenhoven & Broeders (1997); SSBE vowels based on Carley et al. (2018); Cruttenden (2008); Roach (1991); Wells (1982b).

The /ʌ/-/ɒ/ vowel contrast is less often discussed in this context, but is no less troublesome for Dutch learners of SSBE; see Vowel diagram 7 above. In general, Dutch learners will substitute their /ɔ/ vowel, which is between a close-mid and open-mid rounded back vowel, for the English /ɒ/, which is much lower. The main problems are with /ʌ/, for which Dutch does not have any close alternatives. Often, learners will substitute either the /ʏ/ vowel or the /ɔ/ vowel as closest alternatives: this is influenced by the spelling with <u> leading to /ʏ/ and <o> leading to /ɔ/,

since this is a consistent mapping in Dutch. Note that while Dutch /ɑ/ would potentially be a closer option in some cases and contexts, this does not seem to be a common substitution; this might be due to its lack of lip-rounding, or because of its consistent mapping to <a> in orthography.



Vowel diagram 8. Commonly confused SSBE /uː/-/ʊ/ vowel contrast, with Dutch closest equivalent vowels. SSBE vowels indicated in red, East Brabantian Dutch closest equivalents or common substitutions indicated in black. East Brabantian Dutch vowels based on Booij (1999); Gussenhoven (2002, 2007); Gussenhoven & Broeders (1997); SSBE vowels based on Carley et al. (2018); Cruttenden (2008); Roach (1991); Wells (1982b).

The /uː/-/ʊ/ contrast in SSBE is one that is difficult for Dutch learners. Generally speaking, they do not perceive the difference between the two vowels, again partially linked to the English spelling. The /uː/ vowel is perhaps closest to the Dutch /u/, as seen in Vowel diagram 8, although it is significantly longer and is currently undergoing fronting so its suitability in terms of mapping may vary depending on the SSBE speaker. The /ʊ/, on the other hand, is unlike any Dutch vowel, and the closest option happens to be /u/ as well. In this case, the vowel

length is similar, but the lowering of the jaw differs. Since both vowels are mapped onto the same Dutch vowel, this effectively reduces the contrast and can lead to confusion. However, the /uː/-/ʊ/ contrast is a less productive contrast in terms of the lexicon, so in practice this will lead to less confusion than the collapsing of the /e/-/æ/ contrast. This may explain why learners are not as aware of this problematic contrast.

Looking at the relative difficulty of these vowel contrasts in light of PAM, the /iː/-/ɔː/ control contrast will result in two category assimilation, making the contrast very easy to distinguish. The other three contrasts will all result in single category assimilation with a potential category goodness difference. The amount of category goodness difference for each of these would determine the relative difficulty; given the descriptions above, /ʌ/-/ɒ/ is likely to be less difficult to tell apart than /e/-/æ/ and /uː/-/ʊ/, which might lead to similar amounts of difficulty.

2.3 Summary

This chapter provided an overview of the key differences between the phonetics of Dutch and English, and in particular those between East Brabantian Dutch and Standard Southern British English as those are the varieties used in the rest of this thesis. The vowels are of particular interest in this case, as they are key to the stimuli used in further experiments.

There are limited resources on the perception and production difficulties of English learners of Dutch. The pilot study run for this thesis showed there are three vowels in Dutch that are often confused by native speakers of English: /øː/-/ɑu/-/œy/. The two individual elements that are part of the /œy/ diphthong have no English equivalent, nor does the currently diphthongising /øː/, making the confusion between these two vowels quite predictable. The confusion between /ɑu/ and /œy/ was more surprising, but can be explained by a sound change in process in English: /u/-fronting means both [u] and [y] are assimilated to the English /u/ phoneme, opening up the possibility of the Dutch /œy/ being interpreted as a backing rather than a fronting diphthong, and causing both /ɑu/ and /œy/ to be perceived as variants of the same SSBE vowel. Another explanation lies in exposure to regional variation of the MOUTH vowel, with Northern Irish variants being particularly close to the Dutch /œy/.

A much more extensive literature focussing on Dutch learners' of English suggests that despite extensive media exposure, Dutch native speakers still come up against several difficulties when learning SSBE vowels. The /e/-/æ/ contrast is predicted to be difficult for Dutch learners as they have no close equivalent for /æ/ and, therefore, neutralise the contrast by collapsing both English vowels onto their Dutch /ɛ/. The contrast between /ʌ/ and /ɒ/ is predicted to lead to similar difficulty

due to a lack of an equivalent for /ʌ/, but with the added influence of native orthography leading to a predictable substitution pattern. Finally, the /u:/-/ʊ/ contrast is predicted to lead to problems for Dutch learners as both vowels are equally dissimilar to the Dutch vowel /u/, which is the closest approximate onto which they could map.

This overview of the phonetics of English and Dutch justified the choice of vowel contrasts used in the training stimuli in the experiments reported in Chapter 3 (English learners of Dutch vowels) and Chapter 4 (Dutch learners of English vowels) respectively. The aim was to use vowel contrasts where the participants would have difficulty in distinguishing the vowels at pre-test, to give maximum possibility of seeing changes due to training at post-test.

Chapter 3. Study 1

3.1 Introduction

Overall, the little research that has been done to directly compare HV versus LV input seems to be unable to provide an unequivocal answer as to how input variability affects the acquisition of non-native speech sounds. This is particularly true for child learners, as the two studies conducted with children (Evans & Martín-Alvarez, 2016; Giannakopoulou et al., 2017) found conflicting results. While both found that children improved more quickly with single-talker input during training, the results at test differed. Evans & Martín-Alvarez (2016) found the predicted benefit of multiple talker input (HV training) at test, although only when both talkers and items tested were novel. Giannakopoulou et al. (2017) found an unpredicted benefit of single-talker input (LV training) for all test items. Both studies found an advantage of LV input in training. One other relevant study with children is reported in the literature, comparing multiple talker input for L2 vocabulary learning (Sinkevičiūtė et al., 2019). This study found a benefit of HV training at test for adults, in line with other literature, but *not* for children, with Bayes Factor analyses finding evidence for the null.

The present study further investigates whether the HVPT paradigm is useful for L2 speech sound learning in children and, if so, whether they show the same HV advantage as adults have been thought to do (Lively et al., 1993). This study used

participants who were naïve learners of the L2. While this approach differs from many L2 training studies, it has the advantage of ensuring that the two age groups have matched exposure to the non-native contrast (i.e. none).

In this case, native English speaking adults and children were taught Dutch vowels in a single training session, receiving either HV (four talkers) or LV input (one talker). This study tested 7-8 year old children, in line with the ages tested in Evans & Martín-Alvarez (2016); Giannakopoulou et al. (2017); Sinkevičiūtė et al. (2019), and adults for comparison purposes. In the HV condition, talker variability in the training input was blocked (one block of training for each of the four talkers) mitigating difficulties in continuous switching between talkers, following Perrachione et al. (2011).

The studies by Evans & Martín-Alvarez (2016) and Giannakopoulou et al. (2017) both primarily assessed generalisation using a discrimination task. The current study also included both a pre-/post-test discrimination task, allowing comparison to these studies. Importantly however, it also included an identification task at post-test, bearing in mind that Sadakata & McQueen (2013) found the HV advantage only showed up in identification but not discrimination. Identification tests generally use orthography (click on "r" or "l"), allowing both trained and untrained words to be included. However, this was not possible here, since all participants in this study were naïve learners with no experience of Dutch orthography, and thus do not have knowledge of which L2 vowel phonemes correspond to the Dutch spelling. Instead, they are likely to activate L1 vowels that

map onto the same spelling. In this case, the Dutch orthography maps onto existing English vowels, which could lead to unwanted transfer from English as the English vowels are quite different from the Dutch targets. Consequently, the aim was to adapt the paradigm in such a way to make it less reliant on orthography while still being able to use a similar type of identification test. Thus, Experiment 1 investigated the effect of talker variability through means of a HVPT study that used geometrical shapes as an approximation of orthography. The same task was used in training as in the identification task: a 3AFC task where the participant hears a word, and selects a geometric shape, depending upon the vowel in the word (similar to Wang & Kuhl (2003) where animals were associated with specific tones being trained). A key goal is to investigate whether exposure to multiple talkers aids generalisation to novel talkers, and thus all test items in both the Discrimination and Identification task use untrained talkers. In addition, both trained and untrained *items* (words) were included in both tests. It is possible that talker variability might promote generalisation across other dimensions; variability encountered from hearing different talkers – i.e. the idiosyncratic variations in how they pronounce particular phonemes – might be related to the type of variability that is needed to recognise phonemes across different acoustic contexts and phonetic environments. This could explain why Evans & Martín-Alvarez (2016) only found a generalisation benefit in the HV condition, and in general the HV benefit might be expected to be stronger for untrained items than for trained items.

In terms of the vowel contrast which was targeted, this study chose to focus on the Dutch 3-way contrast /øː/-/ɑu/-/œy/, since this contrast is known to be difficult for native English speakers (as discussed in Section 2.1.2). Training and test items were chosen with respect to two considerations: first, stimuli were aimed to vary in their phonetic environment (i.e. for the vowels to be surrounded by a range of consonants), and second, the aim was for participants to be trained on real Dutch words (rather than non-words) wherever possible. This was not only done to increase the ecological validity of the study, but it was also an important part of recruitment: teaching children a real language means schools are more likely to participate (note that this is particularly relevant in the later experiments where participants are trained to associated words with meaningful pictures).

The key hypothesis for Experiment 1 is that there should be a benefit of having had HV input in training on generalisation to novel speakers and items. This would show up as greater performance of participants in the HV condition than LV condition in the Identification task at post-test (main effect of variability condition), and greater improvement for that group from pre- to post-test in the Discrimination task (interaction between variability condition and test session). Following Evans & Martín-Alvarez (2016), it is possible that this HV benefit may be greater for untrained items than trained items (interaction between item novelty and variability condition in Identification, and interaction between item novelty, variability condition, and test session in Discrimination). For children, results from Giannakopoulou et al. (2017) suggest there could alternatively be an

LV benefit at test. Therefore, for child participants only, this experiment tests both for an HV benefit as well as an LV benefit in the Discrimination and Identification task. In addition to looking at performance at test, performance across the four blocks of training will also be investigated. Following Evans & Martín-Alvarez (2016) and Giannakopoulou et al. (2017), faster learning is expected in the LV condition, which should show up as an interaction between block and variability condition.

3.2 Experiment 1

3.2.1 Method

Participants

Participants were 48 children[1] (mean age = 8;2 years, SD = 4 months) recruited from a primary school in North London, and 48 adult (mean age = 26;4 years, SD = 8;9 years) native speakers of English recruited through the UCL psychology subject pool. Several participants reported speaking more languages than just English[2]. For

[1] 2 additional children were tested but were not included in the analysis as they did not complete the full experimental session due to technical failure.

[2] 9 adults reported only speaking English, 26 additionally spoke French, 13 Spanish, 10 German, 8 Mandarin, 3 Malay, 3 Hindi, 2 Bengali, 2 Cantonese, 2 Yoruba, 1 Poshwari, 1 Japanese, 1 Korean, 1 Portuguese, and 1 Danish. All children were learning French at school. 31 children reported speaking no other languages, 5 additionally spoke Ghanaian, 3 Romanian, 2 Spanish, 1 Urdu, 1 Bulgarian, 1 Turkish, 1 Hungarian, 1 Polish, 1 Shona, 1 Yoruba, 1 German, 1 Arabic, and 1 Greek.

these languages, it was checked whether they used the vowels that were part of training; this was not the case for any of the languages, and therefore, participants were not excluded. A separate set of 6 native speakers of Dutch (mean age = 49;2 years, SD = 24;3 years) was recruited as a control group for the pre-test.

All participants had normal or corrected-to-normal vision, unimpaired hearing, and none were dyslexic or had a language impairment. Children were tested individually by a researcher in their school, adults were tested in a sound-attenuated booth at UCL, and Dutch native speakers were tested in a quiet room in the Netherlands. For all children, informed opt-in consent was obtained from a parent or caretaker prior to testing, while adults signed a consent form before participating. Participants were randomly assigned to one of the counterbalanced versions of the two experimental conditions (three in each of the eight versions of the HV and LV conditions – see Table 4). In return for their participation, children received stickers during the experiment and a certificate after completing the session, while adults received either 0.5 credits or a payment of £4.

Stimuli

Stimuli consisted of 12 monosyllabic minimal triplets, all CVC-contexts containing a Dutch vowel (see Table 3). Four three-way pairs were used in both training and each of the tests, four novel pairs were used in discrimination (pre- and post) only, and four pairs were used as novel items in post-test identification only.

Trained items	Novel items for Discrimination	Novel items for Identification
faun – *föhn* – fuin /fɑun/ /fø:n/ /fœyn/	saum – seum – suim /sɑum/ /sø:m/ /sœym/	sauk – seuk – suik /sɑuk/ /sø:k/ /sœyk/
koud – keut – *kuit* /kɑut/ /kø:t/ /kœyt/	*pauk* – *peuk* – *puik* /pɑuk/ /pø:k/ /pœyk/	taup – teup – tuip /tɑup/ /tø:p/ /tœyp/
mauf – meuf – muif /mɑuf/ /mø:f/ /mœyf/	kaum – keum – kuim /kɑum/ /kø:m/ /kœym/	*mout* – *meut* – *muit* /mɑut/ /mø:t/ /mœyt/
naus – *neus* – nuis /nɑus/ /nø:s/ /nœys/	mauk – *meuk* – muik /mɑuk/ /mø:k/ /mœyk/	*kous* – *keus* – *kuis* /kɑus/ /kø:s/ /kœys/

Table 3. List of stimuli used in Experiment 1. The words in italics are real Dutch words.

For each minimal pair, the consonant context was kept constant and the critical difference was found in the vowel only. The three Dutch vowels used were <au> /ɑu/, <eu> /ø:/, and <ui> /œy/; these vowels were chosen as they were shown to be particularly difficult to discriminate for English speakers in a pilot task (as discussed in Chapter 2). Stimuli were visually represented by three geometrical shapes that were consistently mapped to the vowel used in the stimulus: a blue shape to /ɑu/, a red shape to /ø:/, and a yellow shape to /œy/.

Consonant contexts for the stimuli were phonetically legal in Dutch; some resulted in real Dutch words, while most were pseudo-words. The consonant contexts were created according to the following constraints: (1) Approximants (/r/, /l/, /w/) were not included as they have a significant influence on the vowel quality (Collins & Mees, 2003) and show a lot of pronunciation variation between regions and even within speakers of Dutch. (2) Marginal Dutch consonants (/ʒ/ and /g/) as well as

non-English consonants (/x/, /ɣ/, /χ/, /ɦ/) were not included to avoid confusion. (3) Nasal /ŋ/ was disregarded due to its inability to be preceded by a long vowel or diphthong. (4) Only voiceless plosives and fricatives were taken into consideration, as Dutch has final devoicing, and partial word-initial devoicing (Gussenhoven & Broeders, 1997). (5) The contexts varied in place of articulation across the items (creating allophonic variation); see Table 3 for the full list of items. To ensure the contexts were equally balanced across trained and novel items, the place of articulation of both the onset and coda consonants in the novel items was closely matched to those of the trained items.

Some stimuli recordings were made in a sound-attenuated booth at UCL and others in a quiet room in the Netherlands at a sampling rate of 44 100 16-bit samples per second. They were later downsampled to 22 050 samples per second, had their amplitudes equalized to 70 dB, and were filtered using Praat (Boersma & Weenink, 2015). The stimuli were recorded by six native speakers (4 female, 2 male) of a southern variety of Dutch (East Brabantian). The words were read multiple times in various random orders, and the two most neutral recordings for each word per speaker were selected so as to have multiple instances of the same token. Final stimuli can be found on the OSF at https://osf.io/wprs5/.

Design

Each participant completed one experimental session, which involved three stages: pre-test, training, and post-test. The pre-test consisted of a discrimination task and

a short stimuli introduction where participants were first exposed to the mapping between spoken words and the shapes, which was followed by four blocks of training. The post-test started with a discrimination task identical to the pre-test, followed by an identification task. During training and in the identification task, participants were asked to map the stimuli to one of the three shapes on the screen.

There were two experimental conditions: high variability (HV) and low variability (LV) training. The conditions only differed in training, where in the HV condition participants heard four different speakers (2 male, 2 female), while in the LV condition stimuli were spoken by a single talker (one of the four used in HV). The four speakers in HV were organised in blocks so that participants did not have to adapt on a trial-by-trial basis, since this has been shown to be detrimental (Perrachione et al., 2011). To match the training task across conditions, LV training was also organised in blocks, but participants heard the same speaker across all four blocks. Pre- and post-tests were identical across conditions, and all stimuli in these tests involved a novel speaker who had not been encountered during training.

For each condition, the use of different speaker in different tasks was counterbalanced across participants. In training, the order of the four speakers was rotated across versions so that each speaker occurred in every block across the different versions, to counterbalance against any chance differences in intelligibility across speakers. The speaker for the shape introduction was always the first speaker used in training. The discrimination task and identification task in the pre- and post-test included two new female speakers not used in either of the

training conditions. These two speakers were rotated between the discrimination and identification task to once more avoid any confound of speaker intelligibility. The counterbalancing resulted in 8 counterbalanced versions per condition, so 16 versions in total (see Table 4 for the full counterbalanced versions).

	v1	v2	v3	v4	v5	v6	v7	v8
Discrimination	F1	F2	F1	F2	F1	F2	F1	F2
Shape introduction	F3	F3	F4	F4	M1	M1	M2	M2
Training*	F3	F3	F4	F4	M1	M1	M2	M2
	F4	F4	M1	M1	M2	M2	F3	F3
	M1	M1	M2	M2	F3	F3	F4	F4
	M2	M2	F3	F3	F4	F4	M1	M1
Identification	F2	F1	F2	F1	F2	F1	F2	F1

Table 4. Counterbalanced versions of both low variability (LV) and high variability (HV) conditions of the experiment. 'F' indicates a female speaker, 'M' indicates a male speaker. *Note that all four training speakers occur in HV, but that for LV the first training speaker is used throughout all four blocks.

Procedure

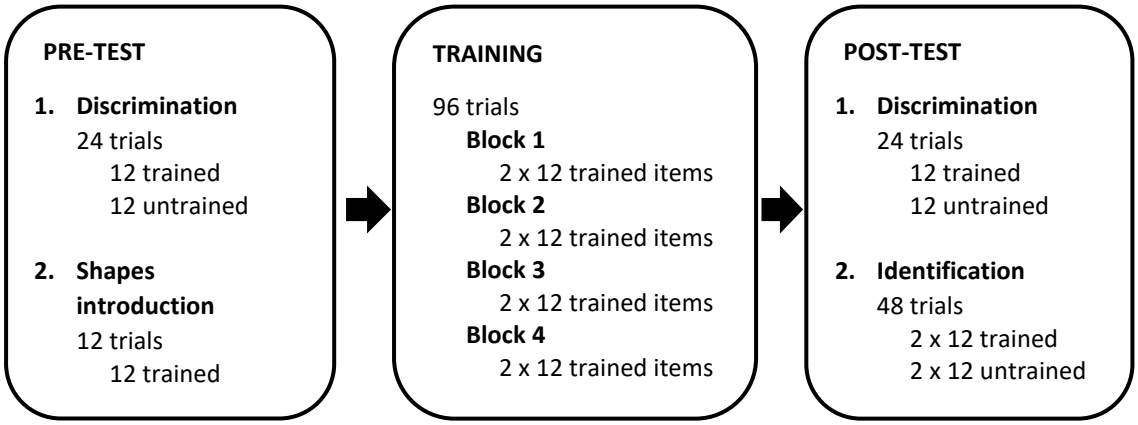


Figure 1. Number and type of experimental trials per task for experiment 1.

Figure 1 shows the number of experimental trials per task; the procedure for each task is described below. All tasks were run using PsychoPy (Peirce, 2007) on a laptop computer in quiet classrooms at the school or in a sound-attenuated booth at UCL. The native Dutch speakers only performed the pre-test discrimination task.

Stimuli were presented binaurally over headphones at a comfortable listening level. The experimental program recorded participants' response accuracy.

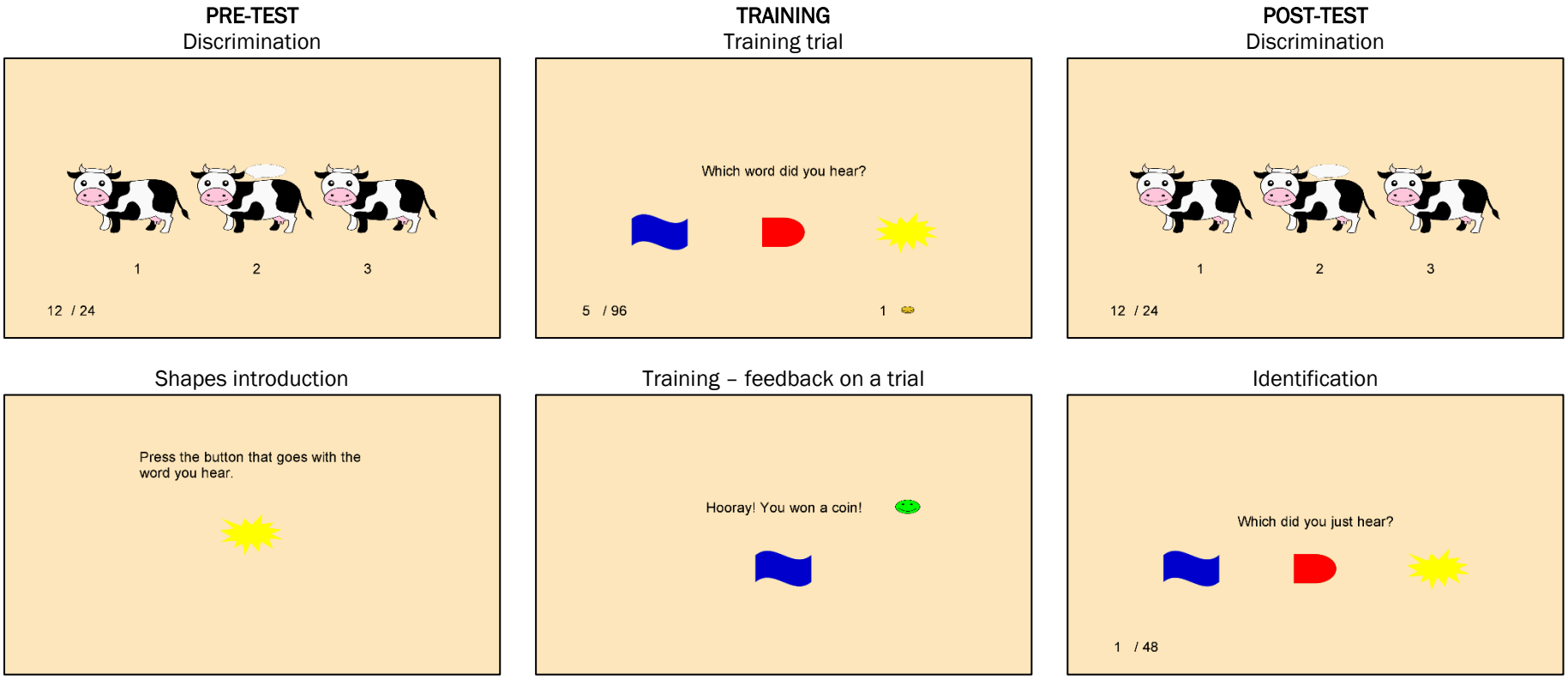


Figure 2. Screenshots from each of the experimental tasks of Experiment 1.

Discrimination task

The discrimination task was identical pre- and post-training, and was a 3-alternative forced-choice (3AFC) oddity task. In the task, three clipart pictures of a cow appeared on the screen (see Figure 2). Participants heard three words played (ISI 200ms), two of which were different tokens of the same word (e.g. '*mauf*', '*mauf*'), and one of which was a different word but always part of a minimal pair with the other two (e.g. '*meuf*'). Participants indicated which of the three words was the odd one out by pressing the corresponding numerical keys (1,2,3) on the keyboard. All tokens were spoken by a single speaker who was not used in training in order to test generalisation across talkers (see Table 4). The task contained 12 pairs which also occurred in training (trained items), and 12 pairs which were novel (untrained items) occurring once each in random order. Note that all test items probe generalization to a novel talker and the novel items tested generalisation abilities across items. The order of the tokens was fully randomised, and all possible pair combinations occurred, so that all vowels occurred equally as target and foils. No feedback was provided, and the program did not allow for a response before the third word had finished playing.

Shapes introduction

This task was included to familiarise participants with the shapes and sounds that would occur in training. In this task, participants heard the 12 items they were to be trained on once each in random order, spoken by the same speaker as used in the first block of training. For each trial, they heard the word, saw the

corresponding shape and were asked to press the corresponding button from one of three keys on their keyboard (each marked with a sticker for one of the three possible shapes). The three shapes were chosen to not be easily nameable, were the same throughout the experiment, and always corresponded to the same vowel: a blue shape to /ɑu/, a red shape to /ø:/, and a yellow shape to /œy/ (see Figure 2).

Training

Training consisted of four blocks with a break between blocks. Blocks occurred in a fixed order with trial order randomised across blocks. On each trial, participants heard a word and selected one of three candidate shapes displayed on the screen (see Figure 2) by pressing its respective button. In each block, all 12 trained items were repeated once, resulting in 24 trials per block, 96 trials in total. For LV, all tokens were spoken by the same speaker while for HV, speaker changed per block.

If participants selected the correct shape corresponding to the item they heard, a green happy face was presented as well as the English text 'Hooray! You won a coin!' (see Figure 2). Simultaneously, the correct word was repeated and the correct shape was shown. Additionally, the number of coins earned (indicated by a number in the bottom-right corner of the screen) was incremented. If an incorrect shape was selected, a red sad face was presented as well as the text 'Too bad, better luck next time!'. Again, the correct word and shape were repeated. The total number of coins a participant had won was always visible during the training trials (though not during feedback), as was the total number of trials. At the end of training, participants were shown a screen that indicated the total number of coins

they had earned. The experimental program recorded participants' trial-by-trial accuracy.

Identification task

This task was identical to the training task except that (a) no feedback was given (and no coins were received), and (b) 12 untrained items were included on top of the 12 trained items. The untrained items differed from those occurring in discrimination. All items, both trained and untrained, were spoken by a new speaker, and were repeated once. The experimental program recorded response accuracy (see Figure 2 for an example of a trial from the identification task).

Analyses

Data were analysed using generalised logistic mixed effects models (see Baayen, Davidson, & Bates (2008)) using the lme4 package (Bates, Mächler, Bolker, & Walker, 2015) for the R environment (R Core Team, 2018). The *knitr* (Xie, 2018) and *kableExtra* packages (Zhu, 2018) were used to present the results. Graphical representations were created using *ggplot2* (Wickham, 2016) and *ggpirate* (Braginsky, 2018).

Adult and child data for each task were analysed separately. The approach to model building was to include all experimentally manipulated variables and their interactions for each task as fixed factors in the models (i.e. Discrimination: *session* (pre/post), *condition* (HV/LV), *item novelty* (trained/novel); Training: *block* (1-4), *condition* (HV/LV); Identification: *condition* (HV/LV), *item novelty* (trained/novel)) regardless of whether they contributed to the model, and to

include *participant* as a random effect with a full random slope structure as recommended by Barr, Levy, Scheepers, & Tily (2013). For items, random intercepts were automatically included, however by-item slopes were only included if they contributed to the model following the approach recommended in Matuschek, Kliegl, Vasishth, Baayen, & Bates (2017). The final model structure detailing the random effects structure and included control variables is reported for each model. In all models, predicting variables were centred to reduce collinearity effects between main effects and interactions, so that the intercept corresponded to the grand-mean (in log-odds space) and main effects were evaluated as average effects over all levels. In each case, variables were coded so that a positive fixed coefficient represented as follows: an increase over blocks in Training; an increase from pre- to post-test in Discrimination and Identification; trained items being rated higher than untrained items in Discrimination and Identification; greater performance in HV than LV in Discrimination and Identification at test, as well as in Training (though note that the latter is done to keep coding consistent as greater performance in LV is actually predicted for Training). To assess whether performance was above chance, the intercept was compared to chance. All reported models converged using the Bound Optimization by Quadratic Approximation (BOBYQA optimization, Powell (2009)).

Hypotheses and Inference criteria

Although the statistical models included all of the experimentally manipulated variables and their interactions, the approach in this thesis is to only report and

interpret the fixed effects relevant to the hypotheses. These are laid out in Table 5. It can be seen that these hypotheses include the key hypotheses relating to the effect of input variability, which were described in Section 3.1 above. In addition, these experiments test hypotheses corresponding to the expectation that participants will be above chance overall in each of the tests, will show greater improvement with trained than untrained items for pre- to post-test in discrimination, and will show higher performance with trained than untrained items in the post-test identification task. All of these hypotheses are directional, with their predicted direction based on the existing literature discussed above. Phrasing them in this directional manner means the Bayes Factor analysis described in more detail below can provide evidence for or against the specific predictions, while non-directional phrasing would only allow for testing for the presence or absence of a relationship. With non-directional hypotheses, the analyses alone would not indicate whether any effects go in the predicted direction or the opposite direction, and follow-ups would be required to break down the directionality of the effect.

Experiment	Task	Hypotheses
Experiments 1, 2 & 3	Discrimination (pre/post-test)	Above chance performance Improvement from pre- to post-test Greater improvement for trained items Greater improvement after HV training For children, greater improvement after LV training (not tested for adults) Greater improvement after HV training for untrained items specifically
Experiments 1, 2 & 3	Training (one session, four blocks)	Above chance performance Improvement across blocks Greater improvement in LV training
Experiment 1	Identification (post-test only)	Above chance performance Greater improvement for trained items Greater improvement after HV training For children, greater improvement after LV training (not tested for adults) Greater improvement after HV training for untrained items specifically
Experiments 2 & 3	Identification (post-test only)	*For minimal pairs and non-minimal pairs separately*: Above chance performance Greater improvement after HV training For children, greater improvement after LV training (not tested for adults)

Table 5. Hypotheses for Experiment 1, 2, and 3.

Note that for children only, two hypotheses were investigated: both the hypothesis that there was greater improvement in the HV than the LV conditions (following the previous adult literature) and that there was greater improvement in the LV than HV condition (following the findings of Giannakopoulou et al. (2017)).

Note that age comparisons are not included in the above table, as it is not clear in what direction hypotheses should be. The literature reviewed in Section 1.3.4 above showed that adults often outperform children overall; however, in these studies they generally start out with some knowledge of the L2, which is not the

case here. Moreover, some studies discussed in Section 1.3.4 found steeper learning in children, although the literature here is mixed. The approach taken in the current study was not to start out with a set of directional hypotheses about age, but rather to run models comparing age effects as follow-up analyses to the separate analyses performed for each age group. Specifically, this study compares performance across age groups wherever an effect was found in at least one of the two groups (i.e. to see whether there was substantial evidence for this effect being bigger in one age group than the other).

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. This provides a measure of the strength of evidence for a hypothesis (H1), compared with the null (H0). As discussed in Section 1.5, these differ from p-values which do not provide the option to evaluate the strength for the H0 ($p > .05$ does not tell us that we should increase our confidence in the null hypothesis, despite this common misinterpretation in the literature). Since Bayes Factors might be less familiar, p-values (computed from z-values) are additionally reported for the reader's convenience, but note that the results are interpreted with respect to the Bayes Factors.

Bayes Factors were computed using the method advocated by Dienes (2008, 2014, 2015), modelling H1 as a half normal, testing a one sided prediction in each case (which is appropriate since directional hypotheses are tested). Note that using a half normal rather than a uniform has the advantage of favouring smaller values

(Dienes, 2014), which is appropriate for this data given that effects are typically small in previous literature. This method of computing Bayes Factors requires three numbers: two to summarise the data (an estimate of mean difference and the standard error (SE)), and one as an estimate of *x,* the predicted mean difference under H1. As a model of the data, the beta values and the standard errors (SEs) from the relevant fixed effects in the logistic mixed effects model were used (note this allows for working in log odds space, meeting assumptions of normality).

To set x, in the absence of pilot data using sufficiently similar materials and procedures, rough estimates of the mean difference under H1 (*x*) were computed based on values from within the current data set (following Dienes & McLatchie (2018)). Across the experiments in the current chapter, wherever possible, an equivalent effect was used from a model with an independent sample of participants of the same age-group doing the same task (e.g. for the comparison with chance for adults in the discrimination task of the current experiment, setting *x* to the value of the difference between chance and intercept in the adult Discrimination model of Experiment 2). Where this was not possible, a value from a different age-group was used (e.g. for the effect of block for children in the Training task of the current experiment, setting *x* to the equivalent value for adults

in the current experiment)[3]. Note *x* can only be estimated in this manner where there is some evidence for an equivalent effect elsewhere in the data, with the value for the effect at least in the predicted direction. In cases where this was not possible, estimations of *x* were based on the basis of how higher-level means constrain the maximum difference to be expected for this contrast, and then setting the estimate to be half of this value (as a half normal distribution is used, so that the maximum is equal to approximately 2SD), as explained in Palfi & Dienes (2019). For example, where a main effect is predicted, the estimate of the effect *x* can be set as the difference between the grandmean (the Intercept) and chance. The logic is as follows: the *maximum* difference between two conditions is seen if one condition is at chance and the other is above chance. In this case the difference between performance in the two conditions will be equal to: 2 * (grandmean - *chance*). This provides an estimate of the *maximum* value of *x,* and thus the estimate is equal to half this value i.e. *x* = grandmean – *baseline* (i.e. the intercept). Alternately, for a main effect where chance performance in one cell is implausible, an estimation of the maximum effect size could be computed as the difference between the average observed performance in the weaker cell and ceiling (99% correct – since log odds 100% is not computable). This was used for Discrimination,

[3] Note that for Experiment 1, values for Training and ID task from the later experiments cannot be used (and vice versa) since these move to use a 2AFC task rather than a 3AFC task.

where performance at pre-test was observed to be very high across all the experiments in this study. The value of *x* used to inform H1, as well as how it was computed, is indicated in the relevant tables presenting the statistics, and full details are also given in Appendix III. Where BF results were reported in-text rather than in a results table, the notation $BF_{H(0,x)}$ is used (following advice by Dienes, see https://osf.io/hzcv6/) to denote a Bayes Factor where *x* is the standard deviation (sd) of the half normal used to model H1.

Bayes factors were interpreted using the convention first set out in Jeffreys (1961) and expanded on in Dienes (2014), where a BF larger than 3 is substantial evidence for H1, a BF between 1/3 and 3 is ambiguous (indicating the data is insensitive to test the hypothesis), and BF smaller than 1/3 suggests substantial evidence for the H0. Since there is subjectivity in how the values informing H1 are determined, a "robustness region" was computed and reported for each Bayes factor, showing the range of estimates of H1 (i.e. the value used as the SD of the half normal) for which our data would support the same conclusion, noted as $[x_1, x_2]$ with x_1 being the smallest SD and x_2 the largest SD leading to the same conclusion (as recommended by Dienes (2017)). To compute the ranges, values were tested – in increments of 0.01 - from a difference of 0 from chance to the log-odds score corresponding to the difference between chance and ceiling performance (deemed to be all but one trial correct: i.e. 97.9% (log-odds 3.8223) in Discrimination; 99.5% (log-odds 5.2465) in Training; 99.0% (log-odds 4.5424) in Identification). 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 >3.8223 (discrimination), > 5.2465 (training), or > 4.5424 (identification), except for ranges of values giving evidence for H0, where the maximum is infinity. Robustness Regions should be interpreted bearing in mind larger values of H1 bias evidence for the *null,* whereas smaller values bias in favour of H1.

The data and analysis script can be found on the Open Science Framework at https://osf.io/wprs5/.

3.2.2 Results

Results will be discussed per task. For each task, the results are discussed per individual age group, before discussing the combined results looking at age differences. The computation of the Bayes Factors reported in the tables can be found in Appendix III. Colour coding has been used to help understand the results at a glance: green means the Bayes Factor shows evidence for the H1, yellow means the Bayes Factor is ambiguous, and red means the Bayes Factor shows evidence for the H0. Note that for completeness sake, all effects related to the hypotheses described above will be reported, even though it may not always be sensible to interpret all effects: e.g. it might not make sense to look for effects of variability if no effect of training is found, or performance is not above chance.

Dutch native speakers: Discrimination task

Of the 6 speakers, 5 responded accurately to all items, and 1 made 1 error out of 24 items, scoring 0.958. Mean accuracy was 0.993, SD = 0.017. This establishes the

level of a true ceiling effect, and is equivalent to the value being used to inform the maximum of the scale when estimating the H1 for Bayes Factor computation.

Adult data

Discrimination task

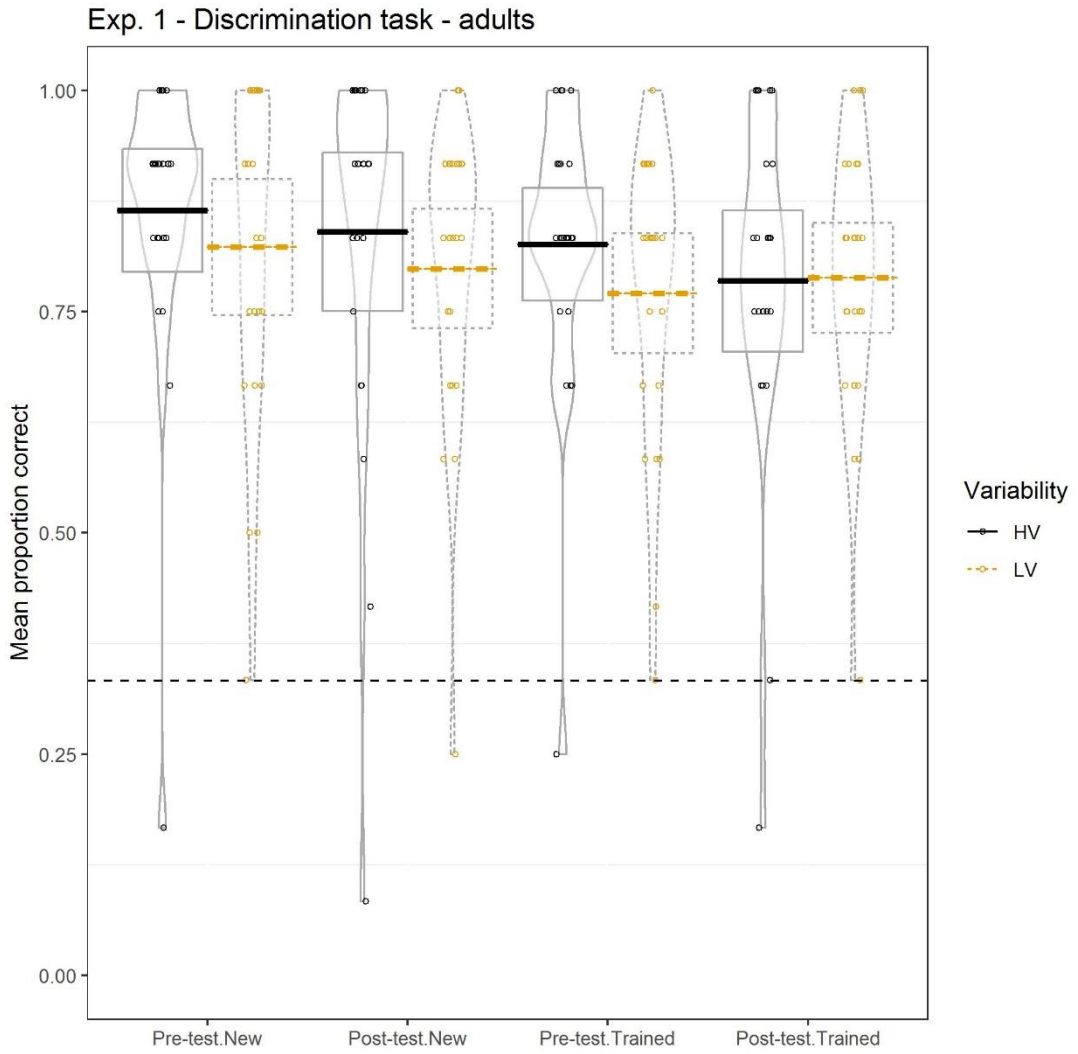


Figure 3. Pirate plot of the accuracy results for adults on the pre- and post-test discrimination task of Experiment 1, comparing accuracy for new versus trained items as well as HV versus LV. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

The data are plotted in Figure 3 and statistics for the key hypotheses tested are shown in Table 6. It can be seen that while there is substantial evidence that participants are above chance on this test (81% compared with 33% chance), there

is substantial evidence that they do *not* improve from pre-training (82%) to post-training (80%). There is also substantial evidence against a difference in improvement for trained and untrained items. Turning to the key hypotheses concerning variability, there is substantial evidence against greater improvement in the HV condition, and against the hypothesis that HV was particularly beneficial for novel items.

Final structure Exp1_Adult_DM model: accuracy ~ session * condition * itemnovelty + speaker + (session*itemnovelty|participant) + (session|contrast)

Hypothesis	fixed effect in model	β	SE	z	P	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept -chance	2.650	0.236	11.231	<.001	2.098	$2.50*10^{26}$	[0.1823, >3.8223]
improve from pre- to post- test	session	-0.201	0.188	-1.074	.283	1.539	0.059	[0.2623, ∞]
greater improvement for trained items	item-novelty : session	0.107	0.305	0.352	.725	1.539	0.131	[0.5723, ∞]
greater overall improvement for HV training	condition: session	-0.188	0.315	-0.596	.551	1.539	0.26	[1.2023, ∞]
greater improvement in HV greater for untrained items	*condition : item-novelty : session	0.426	0.490	0.870	.385	1.539	0.172	[0.7523, ∞]

Table 6. Mixed model results for the Discrimination analysis of Experiment 1, for adults. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Training

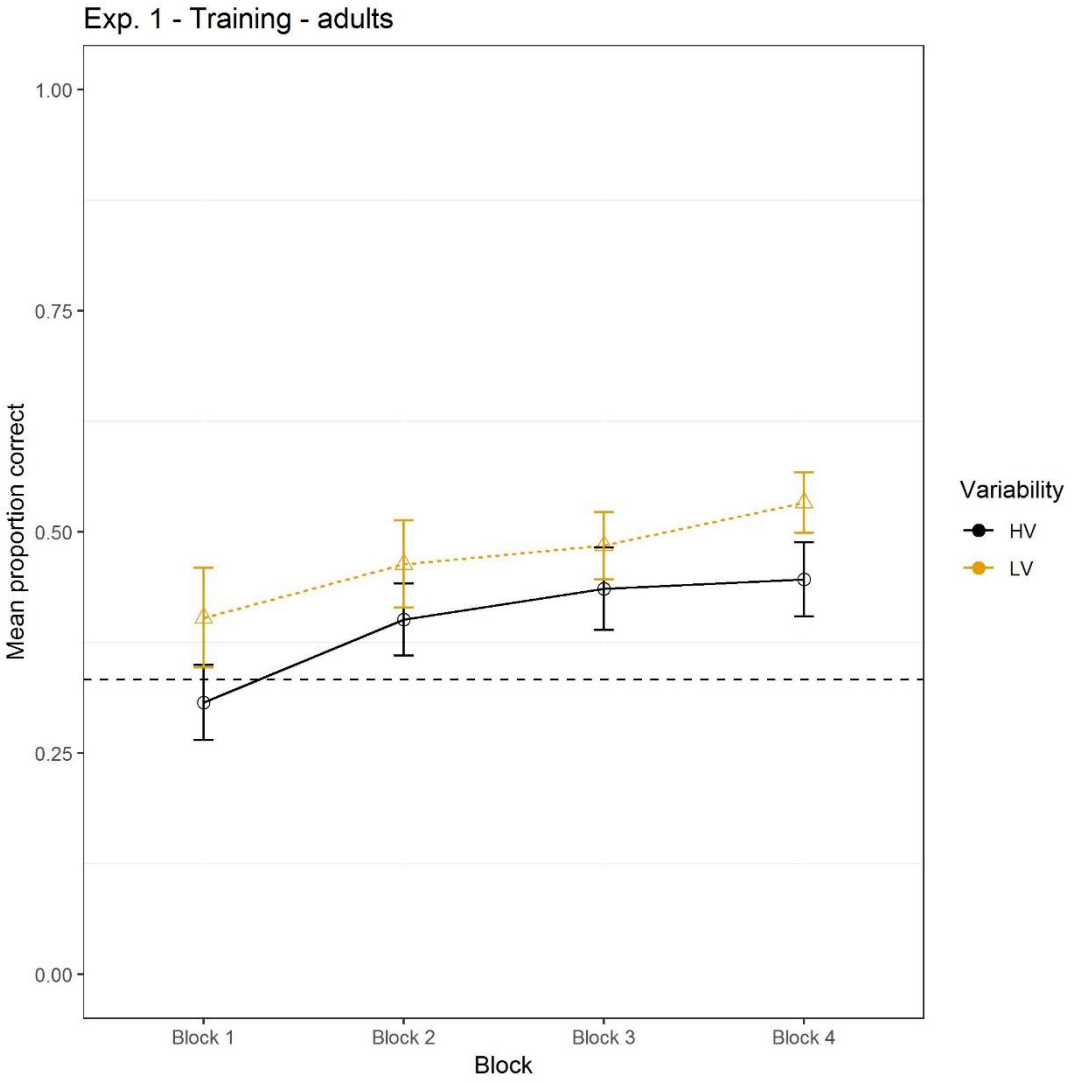


Figure 4. Adult accuracy scores in training of Experiment 1 plotted over the 4 blocks, comparing HV and LV. Error bars show 95% CI, and the dashed line indicates chance level.

See Figure 4 for the plotted data, and Table 7 for statistics for the key hypotheses tested. There was substantial evidence that adults were above chance (43% compared with 33% chance), and improved across training blocks. There was substantial evidence against greater improvement in the LV training condition.

Final structure Exp1_Adult_Train model: accuracy ~ block * condition + speaker + (1|participant) + (condition*block|word)

Hypothesis	fixed effect in model	β	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.418	0.098	4.285	<.001	0.102	166.697	[0.0465, >5.2465]
improvement across blocks	Block	0.189	0.034	5.584	<.001	0.139	$1.17*10^6$	[0.0465, >5.2465]
greater improvement in LV condition	*condition: block	-0.014	0.057	-0.241	.810	0.189	0.242	[0.1365, ∞]

Table 7. Mixed model results for the Training analysis of Experiment 1, for adults. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Figure 5 below shows the degree of variability in the patterns of change over blocks across participants. As can be seen, the majority of adults show a similar pattern of learning across blocks.

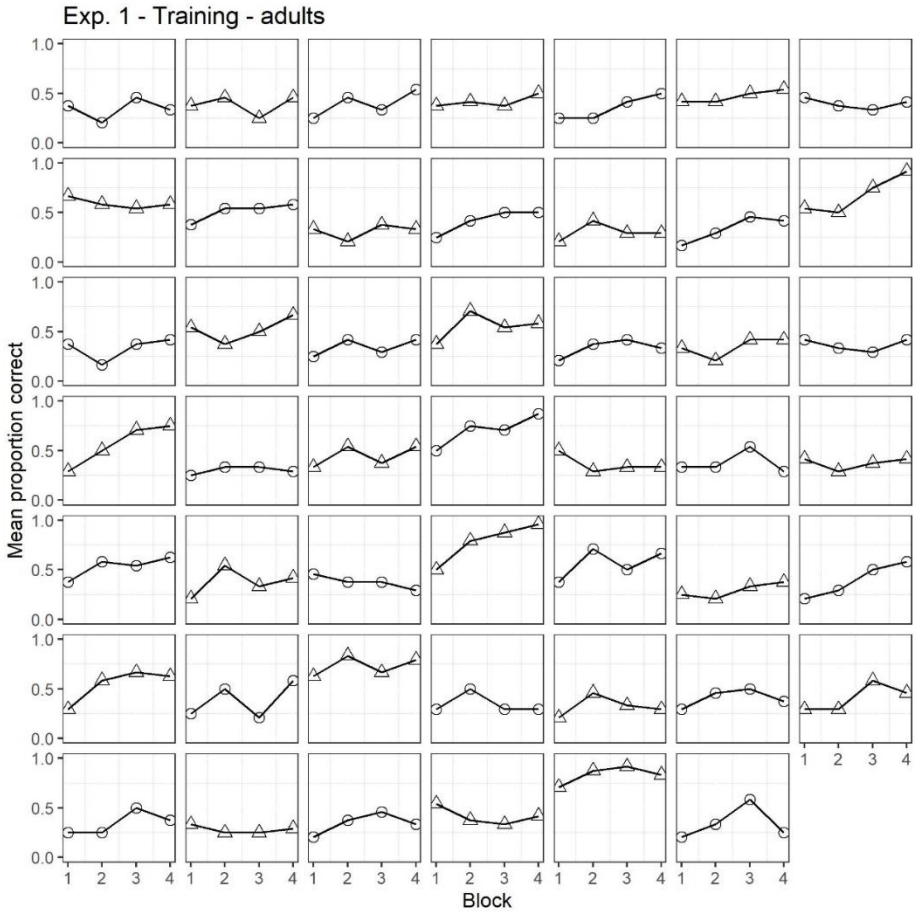


Figure 5. Adult accuracy scores in training of Experiment 1 plotted over the 4 blocks, split out by participant. Plots with circles indicate a participant received HV training, while triangles indicate LV training.

Figure 6 shows the pattern of improvement across the training blocks split out by vowel: while performance on <ui> is lower than that for <au> and <eu>, the pattern of improvement is similar.

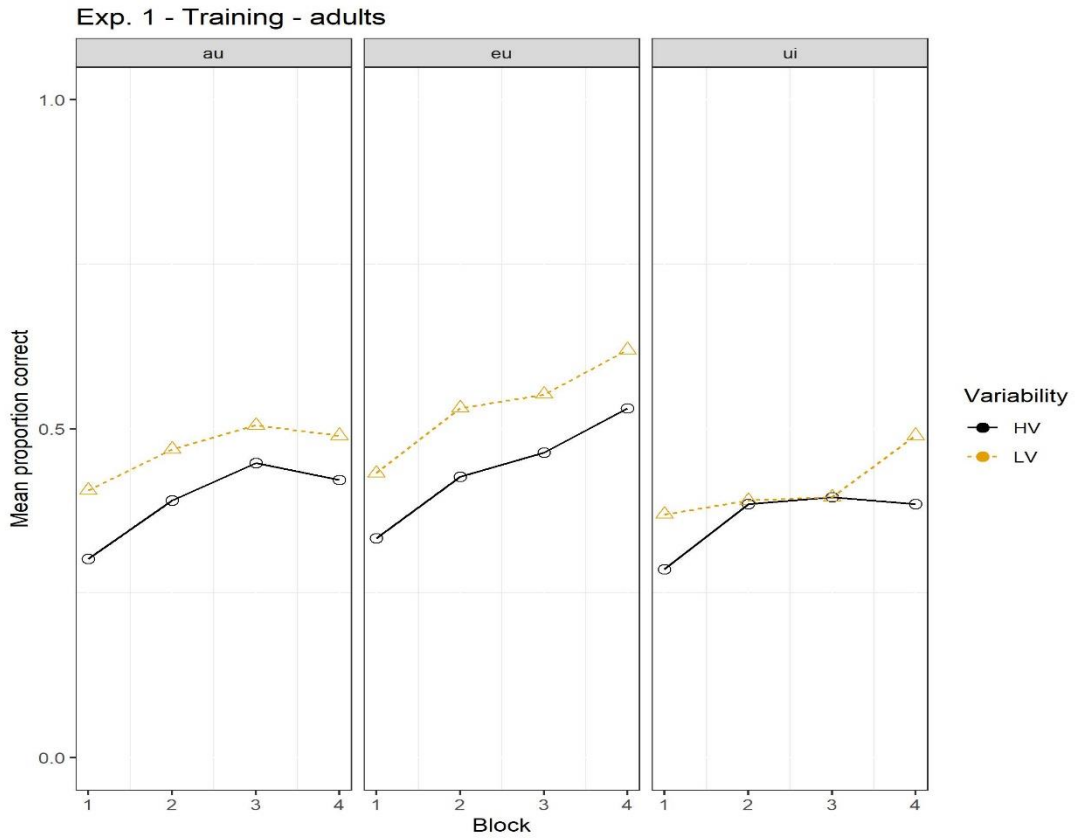


Figure 6. Adult accuracy scores in training of Experiment 1 plotted over the 4 blocks, split out by vowel, comparing HV and LV.

Identification task (post-test only)

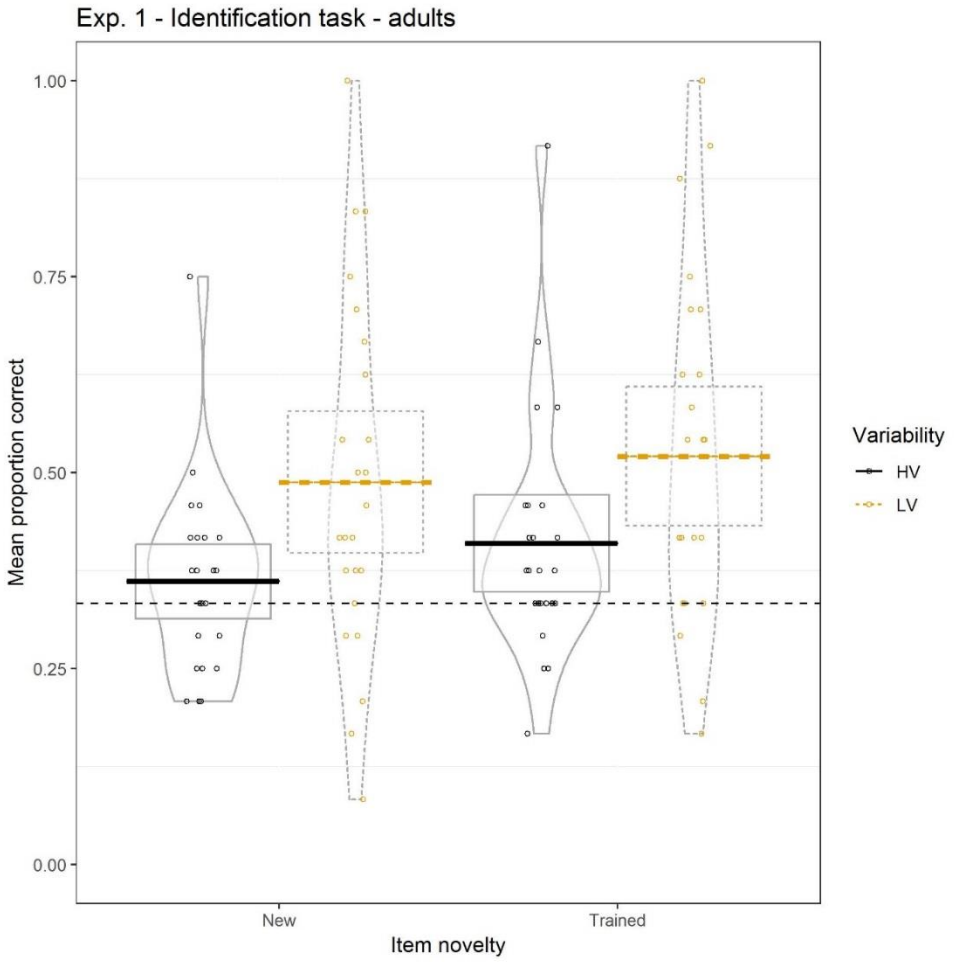


Figure 7. Adult accuracy results on the identification task of Experiment 1, comparing novel and trained items in the two variability conditions. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

There was substantial evidence that adults were more accurate than chance, as seen in Figure 7 and Table 8. The evidence for greater performance with trained items was ambiguous (trained 47%; novel 43%). There was substantial evidence that performance was *not* greater in the HV than LV condition (which recall was predicted given that all test *voices* here are novel; HV condition 39%; LV condition 51%). There was ambiguous evidence for greater performance for novel (compared

to trained) items for participants who had undergone HV (compared with LV) training.

Final structure Exp1_Adult_ID model: accuracy ~ itemnovelty * condition + speaker + (itemnovelty|participant) + (condition|word)

Hypothesis	fixed effect in model	B	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.475	0.124	3.828	<.001	0.418	477.812	[0.0924, >4.5424]
greater performance in HV condition	condition	-0.546	0.231	-2.367	.018	0.418	0.142	[0.1824, ∞]
greater performance in trained than in novel items	Item novelty	0.191	0.129	1.480	.139	0.418	1.346	[0, 2.1224]
greater performance in HV is greater for untrained items	*condition: item-novelty	-0.054	0.183	-0.296	.767	0.317	0.414	[0, 0.4024]

Table 8. Mixed model results for the Identification analysis of Experiment 1, for adults. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Child data

Discrimination task

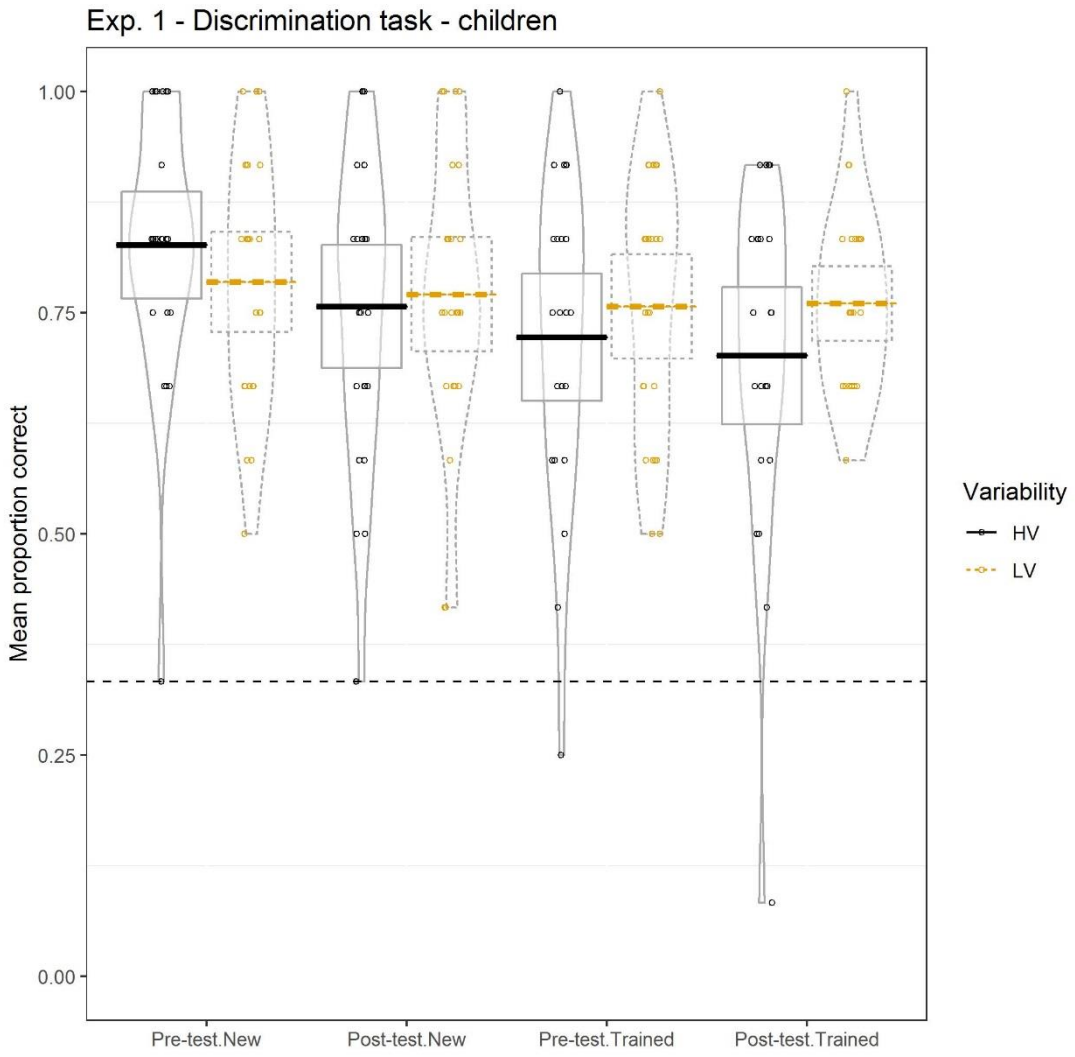


Figure 8. Accuracy results for children on the pre- and post-test discrimination task of Experiment 1, comparing accuracy for new versus trained items as well as HV versus LV. The horizontal line in each violin indicates the mean performance, with the band around showing 95% CI, and the dashed line indicates chance level.

See Figure 8 for the plotted data, and Table 9 for statistics for the key hypotheses.

There was substantial evidence that children were above chance in Discrimination (76% correct compared with 33% chance level). There was no improvement from pre-test to post-test with substantial evidence for the null (pre-test 77%, post-test

75%). There was substantial evidence against the hypothesis that HV would outperform LV, and ambiguous evidence for LV outperforming HV. There was substantial evidence against more improvement in trained than novel items and against the hypothesis that HV training was particularly beneficial for novel items.

Final structure Exp1_Child_DM model: accuracy ~ session * condition * itemnovelty + speaker + (session*itemnovelty|participant) + (session|contrast)

Hypothesis	fixed effect in model	beta	SE	Z	p	predicted effect x	Bayes factor	Robustness Region
above chance performance	intercept-chance	2.029	0.153	13.256	<.001	1.983	1.31 *10^{37}	[0.1823, >3.8223]
improvement from pre- to post- test	session	-0.262	0.131	-1.997	.046	1.665	0.024	[0.1823, ∞]
greater improvement for trained items	item-novelty : session	0.193	0.244	0.794	.427	1.665	0.307	[1.5323, ∞]
greater overall improvement for HV training	condition: session	-0.269	0.227	-1.186	.236	1.665	0.063	[0.2923, ∞]
greater overall improvement for LV training	*condition: session	0.269	0.227	1.186	.236	1.665	0.473	[0, 2.4023]
greater improvement in HV greater for untrained items	*condition : item-novelty : session	0.281	0.420	0.669	.503	1.665	0.302	[0.7223, ∞]

Table 9. Mixed model results for the Discrimination analysis of Experiment 1, for children. *The sign of these fixed effects is changed from the model coding to be in line with the hypothesis.

Training

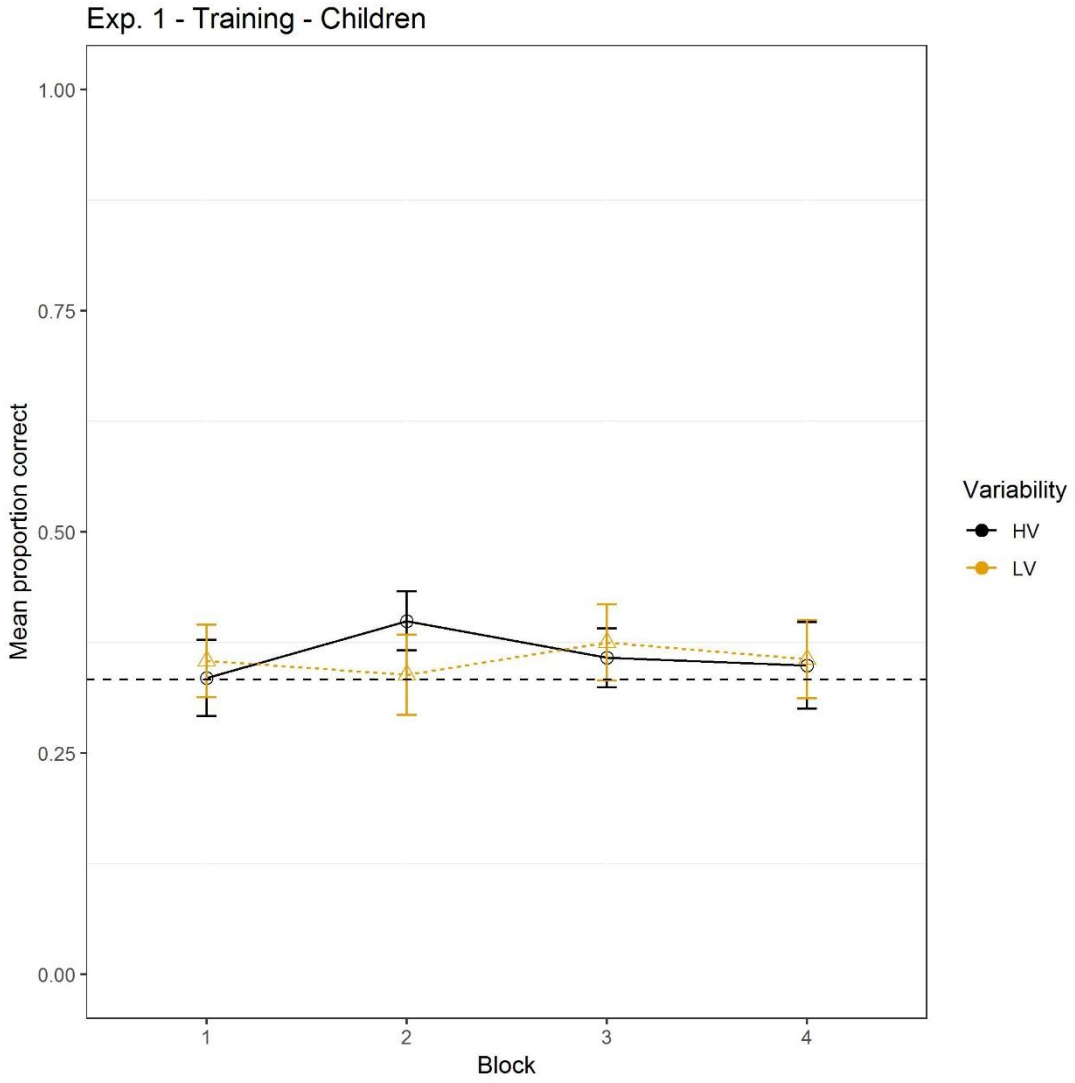


Figure 9. Child accuracy scores in training of Experiment 1 plotted over the 4 blocks, comparing HV and LV. Error bars show 95% CI, and the dashed line indicates chance level.

There was ambiguous evidence for children being above chance in training (36% correct with chance at 33%), see Figure 9 and Table 10. Unlike adults, they did not improve across blocks, with substantial evidence for the null. The evidence for greater improvement in the LV condition was ambiguous.

Final structure Exp1_Child_Train model: accuracy ~ block * condition + speaker + (block*condition||participant) + (block+condition|word)

Hypothesis	fixed effect in glmer model	beta	SE	z	P	predicted effect x	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.102	0.046	2.204	.028	0.418	2.396	[0, 0.3265]
improvement across blocks	Block	0.010	0.033	0.297	.767	0.189	0.221	[0.1265, ∞]
greater improvement in LV condition	*condition: block	-0.017	0.061	-0.280	.779	0.189	0.382	[0, 0.2165]

Table 10. Mixed model results for the Training analysis of Experiment 1, for children.*The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Figure 10 below shows the degree of variability in the patterns of change over blocks across participants. As can be seen, the majority of children show a similar lack of learning across blocks, with relatively stable scores throughout.

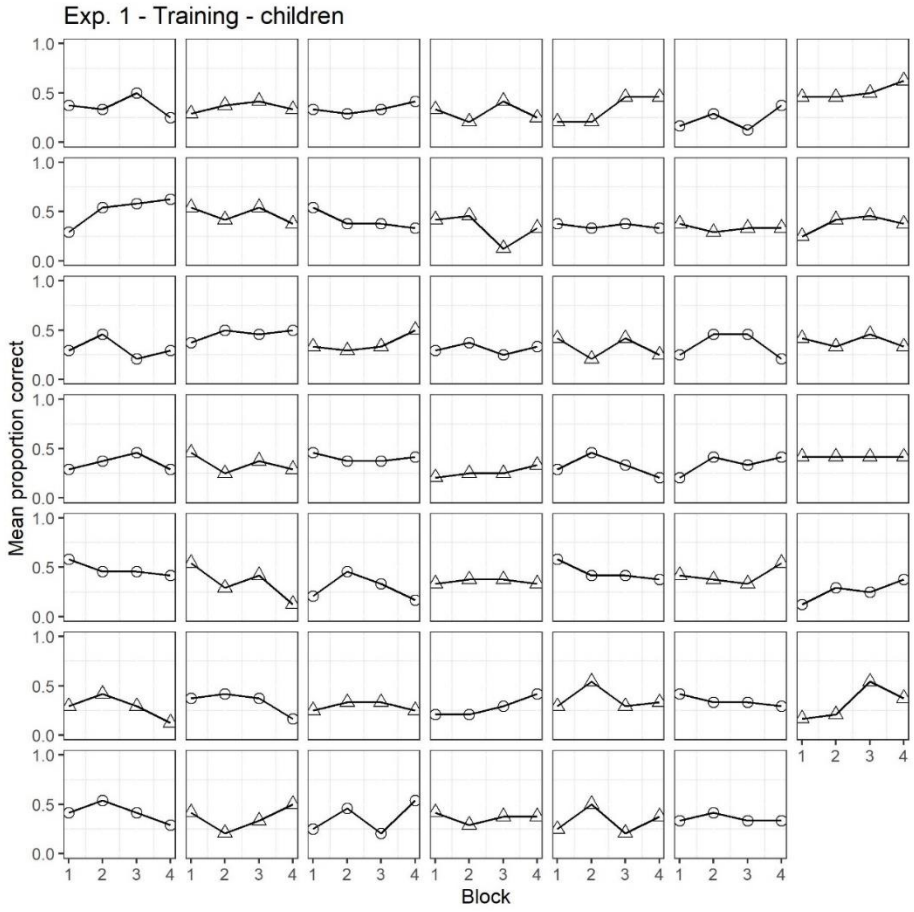


Figure 10. Child accuracy scores in training of Experiment 1 plotted over the 4 blocks, split out by participant. Plots with circles indicate a participant received HV training, while triangles indicate LV training.

Figure 11 shows the lack of improvement across the training blocks split out by vowel: performance is similarly low for all three vowels.

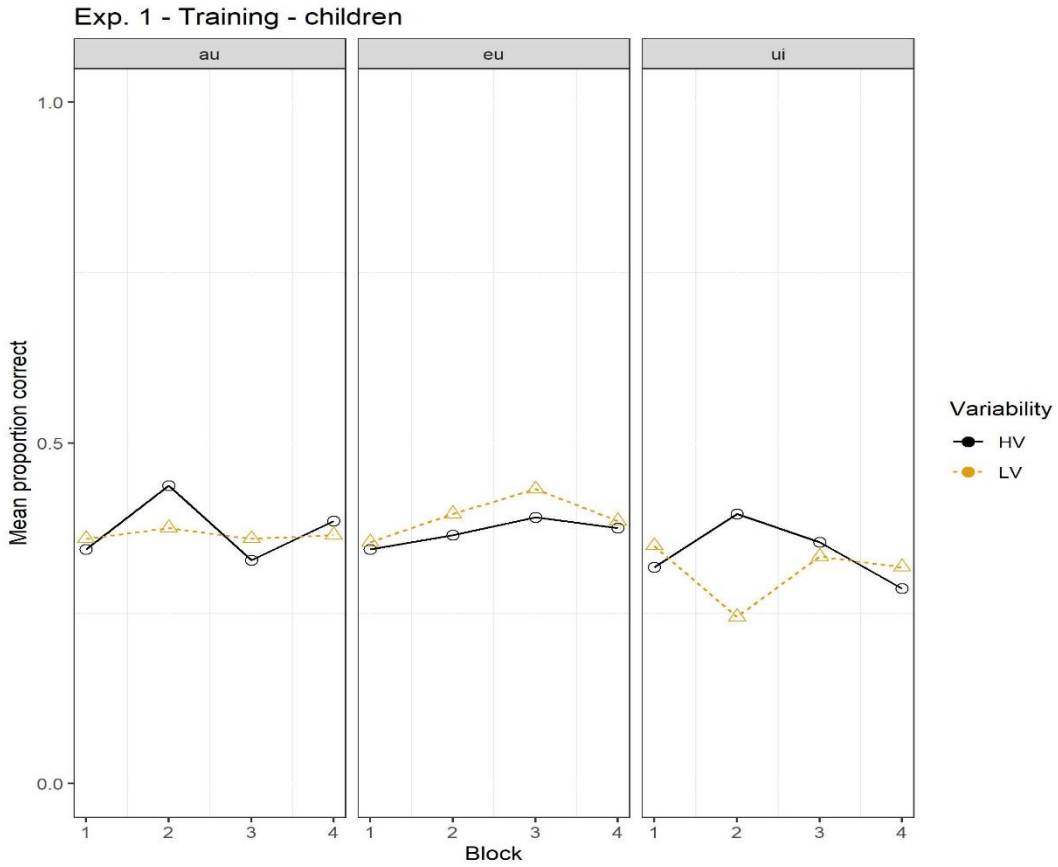


Figure 11. Child accuracy scores in training of Experiment 1 plotted over the 4 blocks, split out by vowel, comparing HV and LV.

Identification task (post-test only)

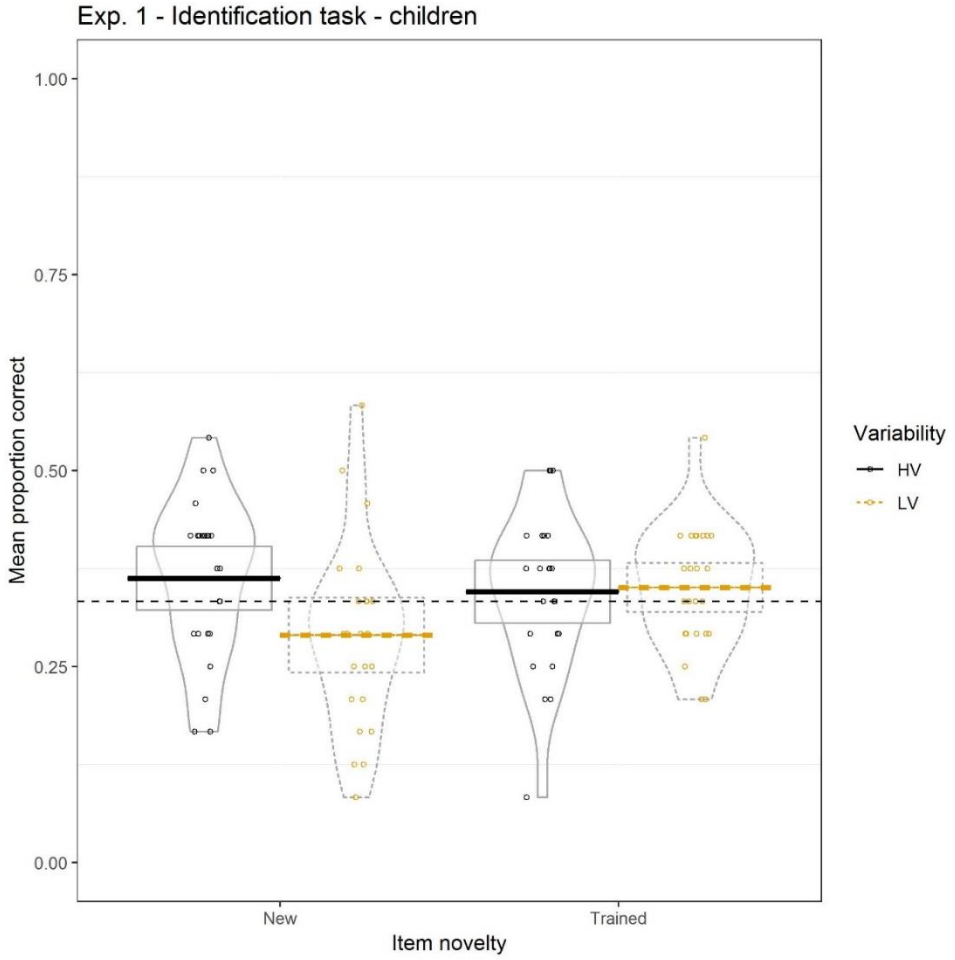


Figure 12. Child accuracy results on the identification task of Experiment 1, comparing novel and trained items in the two variability conditions. The horizontal line in each violin indicates the mean performance, with the band around showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 12 and Table 11, there was substantial evidence against children performing above chance on the Identification task (with performance at 34% with 33% chance level), as well as against LV outperforming HV. The evidence for greater performance in HV than LV condition was ambiguous, as was the evidence for trained items outperforming novel items. There was substantial evidence for the idea that HV training was particularly beneficial for novel items.

Final structure Exp1_Child_ID model: accuracy ~ itemnovelty * condition + speaker + (itemnovelty|participant) + (1|word)

Hypothesis	fixed effect in model	Beta	SE	Z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.004	0.061	-0.065	.948	0.475	0.137	[0.1924, ∞]
greater performance in HV condition	condition	0.159	0.097	1.638	.101	0.475	1.376	[0, 2.1124]
greater performance in LV condition	condition	-0.159	0.097	-1.638	.101	0.475	0.081	[0.1024, ∞]
greater performance in trained than in novel items	Item novelty	0.115	0.121	0.951	.342	0.475	0.622	[0.0924, 0.9324]
greater performance in HV is greater for untrained items	*condition: item-novelty	0.364	0.191	1.909	.056	0.317	3.729	[0.1724, 0.6124]

Table 11. Mixed model results for the Identification analysis of Experiment 1, for children. *The sign of these fixed effects is changed from the model coding to be in line with the hypothesis.

While this interaction between variability condition and item novelty is predicted, overall performance was at chance, and inspecting Figure 12 suggests the interaction might be caused by children (presumably accidentally) performing *below* chance on novel items in the LV condition, while performing at chance in the HV condition. To explore this statistically, each cell was separately tested against chance; see Table 12. This confirmed that there was no evidence that any individual cell was above chance and, critically, that there was no possible evidence for H1 for the novel items in the HV condition.

Hypothesis	fixed effect in model	Beta	SE	Z	p	Robustness Region
above chance performance HV trained	intercept-chance	0.129	0.088	1.466	.143	ambiguous: [0.0924, 1.4324] substantial null: [1.4424, ∞]
above chance performance HV novel	intercept-chance	0.053	0.088	0.605	.545	ambiguous: [0.0924, 0.4424] substantial null: [0.4524, ∞]
above chance performance LV trained	intercept-chance	-0.277	0.190	-1.462	.144	ambiguous: [0.0924, 0.2124] substantial null: [0.2224, ∞]
above chance performance LV novel	intercept-chance	0.077	0.087	0.882	.378	ambiguous: [0.0924, 0.6024] substantial null: [0.6124, ∞]

Table 12. Mixed model results for the follow-up analysis investigating performance compared to chance for the Identification task of Experiment 1, for children. Robustness Regions rather than Bayes Factors are provided as there is no evident value on which to base the estimates.

Comparing adults and children

The analyses reported above suggest some differences between adults and children. Specifically, that adults outperform children in every task, that only adults show evidence of being above chance in Training (while children are ambiguous), improving across blocks in training, and that only adults are above chance in the post-test Identification task. These observations were followed up by combining the data and running the same mixed effect models as above, but this time including age group and – for training only – the interactions *age group by block* and *age group by block by session* as additional predictors. This allowed the use of Bayes Factors to test predictions associated with these hypotheses (note, hypotheses about whether the age groups differ for those effects which were not found in either group were not tested – e.g. improvement for pre- to post-test in

Discrimination, where both age groups showed substantial evidence for the null). Results are in Table 13. In each case there is substantial evidence for the hypothesis that adults outperformed children. This confirmed that adults outperformed children in all tasks, and showed greater improvement across blocks in training.

Final structure Exp1_AgeComp_DM model: accuracy ~ session * condition * itemnovelty + group + speaker + (session*itemnovelty|participant) + (session|contrast)

Final structure Exp1_AgeComp_Train model: accuracy ~ block * condition * group + speaker + (block|participant) + (condition+block|word)

Final structure Exp1_AgeComp_ID model: accuracy ~ itemnovelty * condition * speaker + group + (itemnovelty|participant) + (condition|word)

Hypothesis	fixed effect in model	Beta	SE	Z	P	predicted effect *x*	Bayes factor	Robustness Region
adults outperform 7 years olds in discrimination	Discrimination model: *group*	0.341	0.077	4.407	<.001	1.204	2030.297	[0.1823, >3.8223]
Adults outperform children in training task.	Training model: *group*	0.332	0.044	7.548	<.001	0.257	$3.52*10^{11}$	[0.0465, >5.2465]
Adults show greater increase in performance across blocks that children in training task.	Training model: *group by block*	0.176	0.039	4.521	<.001	0.093	4594.696	[0.0465, >5.2465]
Adults show greater performance than children in identification task.	Identification model: *group*	0.487	0.063	7.754	<.001	0.226	$7.08*10^{11}$	[0.0924, >4.5424]

Table 13. Mixed model results for the Age Comparison of Experiment 1. *The sign of these fixed effects is changed from the model coding to be in line with the hypothesis.

3.2.3 Discussion

In the current study, native English speaking 7-8 year-olds and adults were trained to associate geometrical shapes to 3 Dutch vowels through means of a phonetic training paradigm. The key manipulation was the amount of variability used in the training input, in terms of different speakers producing the stimuli. In addition to measuring performance during training (which used an identity task), participants' generalisation ability was measured in two tasks using untrained speakers: a task measuring their ability to discriminate the vowels before and after training, and a task measuring their ability to identify the vowels associated with the symbols after training. All generalisation items used untrained talkers, while some were trained items and some novel items. The remainder of this discussion will briefly consider how the findings in each task differed from the predictions, before considering the implications of the findings.

Starting with the training task, adults outperformed children and improved more over training. In fact, only adults showed above chance performance and improved over sessions. The evidence as to whether children's performance was above chance was ambiguous but there was evidence that they did *not* improve across training blocks (evidence for the null). In this test, both age groups were predicted to show stronger performance, and potentially greater improvement across blocks, in the LV condition (due to adapting to a single talker). However, there was substantial evidence for the null in adults; since children did not improve across blocks it does not seem sensible to further interpret these effects for this age group.

One reason for not finding this LV effect may be due to the brevity of the training session, as it might be the case that participants need more time in order to adapt to a particular speaker; most studies finding such a benefit have multiple training sessions. This point will be revisited in the general discussion in Section 3.5.

The key hypothesis, however, was that if input variability plays a role in promoting generalisation to novel voices, exposure to HV input would increase performance in the generalisation tests with untrained voices, with this result possibly being greater for untrained items. For children, the alternative hypothesis that they would show greater generalisation following LV training was also tested following Giannakopoulou et al. (2017). However, for discrimination, although both adults and children were above chance in the task, with adults again outperforming children, critically neither age group improved from pre-test to post-test as a result of training, with evidence for the null in each case. This was not modulated by item novelty, training variability condition, or their interaction. In the identification post-test, only adults were above chance, with children at chance with evidence for the null. Critically, regarding differences between conditions, there was substantial evidence that adults did *not* show stronger performance following HV training, and showed ambiguous evidence for having greater performance with trained items. The evidence for an HV benefit specifically in untrained items was ambiguous for adults. These hypotheses are not further evaluated for children, as their performance did not substantially differ from chance.

In sum, while adults showed evidence of learning as a result of the training task, there was little evidence for this in children. Adults showed evidence of generalisation in the Identification task, while children performed at chance. Neither age group showed altered discrimination abilities as a result of training. Critically, no evidence was found of an effect of input variability for either age group in any of the tasks, though note that for children no evidence was found of above chance performance in either the Training or the Identification task in the first place.

Why did Experiment 1 find no evidence of the predicted HV benefit in training for adult generalisation, nor for either an HV or LV training benefit for children? First, it is important to bear in mind that the discrimination task did not seem to be sensitive to learning as a result of training at all, and that child performance on the identification task was at chance throughout. If participants are struggling to learn anything in the task, they might not be expected to show differences between input conditions. Thus it is hard to draw conclusions as to the effect of variability here, at least for children. For adults, in the Identification task, there *is* evidence of learning, yet there is nevertheless substantial evidence that performance is not better in the HV than LV condition. However, inspection of the training data in Figure 4 raises a potential concern: although we have seen that there was no evidence of a steeper slope over block in the LV condition, it appears that the LV condition outperform the HV overall and, critically, even in the first block (where the training has not yet diverged for the two groups). Some follow up analyses

confirm that there is indeed an overall effect of condition in the data (beta = -0.317, SE = 0.158, z = -2.007, p = .045; $BF_{H(0,0.418)}$ = 3.995, RR[0.1256, 0.6665][4]) and that this holds even in block 1 (beta = -0.416, SE = 0.160, z = -2.606, p = .009; $BF_{H(0,0.077)}$ = 2.962, RR [0.0865,3.0765][5]). Thus it seems that the LV participants (accidentally) started out with a higher overall performance than the HV group, and this difference persists throughout training. This raises the possibility that this accidental difference between the participant groups masks a possible HV benefit in generalisation.

Turning to age differences, this experiment did not start with hypotheses about the role of age in this task, since both children and adults start off without prior knowledge of the language. However, age effects observed in the data were followed up with additional analyses. This showed that, while both native English speaking adults and children were able to hear the difference between the three Dutch vowels (both age groups are above chance in the discrimination test at pre-test), there was strong evidence that adults outperformed children in every task. Moreover, only adults showed evidence of being above chance in Training, improving during Training, and generalisation in Identification. Although children's performance in some cases is ambiguous, there is substantial evidence

[4] BF based on the intercept estimate for training.
[5] BF based on the intercept estimate for block 1.

that they are worse than adults in each case, and it seems clear that they are struggling to learn the associations between the non-native vowels and their respective shapes. This is unexpected given previous studies using HVPT in the literature. One possible explanation is the length of training: recall that these types of HVPT studies, whether used with adults or children, usually have more than one training session. For adult participants, studies vary in length between 5 (e.g. in Iverson & Evans (2009); Lengeris & Hazan (2010)) and 15 sessions (in the seminal Lively et al. (1993); Logan et al. (1991)), although the majority of learning seems to take place in the first 10 sessions (Logan & Pruitt, 1995). Typically, studies with children have involved anywhere from 3 sessions (Heeren & Schouten, 2010) to 10 sessions (Giannakopoulou et al., 2017, 2013; Shinohara & Iverson, 2015). The current experiment only consisted of one training session, and the amount of exposure the participants received might not have been sufficient for them to acquire the non-native vowels.

While it seems likely that longer training would aid learning, it is possible that children may have difficulty with the form of the learning task which asks them to master the mapping between the novel geometrical shapes and the abstract vowel categories, designed to mimic orthography. Other than the study by Wang & Kuhl (2003), previous HVPT paradigms that show learning in children have generally either used orthographic symbols with which learners already had some familiarity (Giannakopoulou et al., 2013; Heeren & Schouten, 2010; Shinohara, 2014), or used a picture mapping task where participants mapped whole words to

pictures denoting their meaning (Giannakopoulou et al., 2017; Sinkevičiūtė et al., 2019). In Experiment 1, participants were asked to map symbols to *part* of the word: the vowel phoneme. In some sense this is easier than learning mappings for whole words, as participants only have to learn three consistent grapheme-to-phoneme mappings rather than 36 word-to-picture mappings. However, in order to do this, they have to recognise which part of the word does not change across the different items that map onto the same symbol (e.g. they have to recognise it is the phoneme /œy/ that is stable for all words mapping onto the yellow shape).

Children may find this abstraction process difficult. Word learning biases might make it harder: children might be inclined to assume they are learning labels for "objects", and these should only have one label (Markman & Wachtel, 1988), which makes it hard to associate multiple word stimuli to the same shape.

Since the goal is to find a paradigm where sufficient learning is found in children to be able to see whether HV input is beneficial to learning, in Experiment 2, the paradigm was adjusted to make it more akin to word learning and to try and promote stronger learning for children in Training, by using a picture mapping rather than symbol mapping task in training, similar to Evans & Martín-Alvarez (2016) and Giannakopoulou et al. (2017).

3.3 Experiment 2

Experiment 2 tests English participants learning the same Dutch three-way vowel contrast, this time using picture stimuli rather than abstract shapes, where each picture mapped to the meaning of one of the training words. The key manipulation is once more HV versus LV training input, and generalisation is once more tested with a Discrimination and Identification task. As in Experiment 1, participants are 7-8 year-olds and adults. However, this time two adult groups were recruited: one group was tested in the lab (as in Experiment 1), and one group was tested online. The latter group was recruited in order to make straightforward comparisons with Experiment 3, which used online recruitment. Recruiting the two adult groups also allows for comparison of online and lab-based performance, which has interesting methodological implications, given the concern often expressed that performance will be less good for online participants than in the lab (i.e. due to participants not attending as well, misunderstanding the task etc., see Rodd (2018) for a further discussion of these concerns). Based on this concern, the current experiment included the hypothesis that lab-based participants outperform online participants.

The use of picture stimuli in this experiment meant that in addition to testing whether participants acquired the non-native phoneme contrasts, it was possible to test participants' word learning more broadly. Recall from Section 1.4 that various studies have found phonetically relevant variability can be beneficial in adult L2 vocabulary learning (Barcroft & Sommers, 2005; Sommers & Barcroft,

2011). To date, there is just one study looking at this in child L2 vocabulary learning, and this study did *not* find this predicted HV advantage in children (Sinkevičiūtė et al., 2019). The current study therefore also provides a second investigation as to whether HV training input is beneficial for L2 word learning in children.

3.3.1 Method

Participants

Participants were 48 children[6] (mean age = 8;3 years, SD = 5 months) recruited from a primary school in North London, and two groups of adult participants: 48 adult (mean age = 25;2 years, SD = 10;7 years) native speakers of English recruited through the UCL psychology subject pool, and 48 monolingual native speakers of English[7] located in the UK or Ireland (mean age = 34;2 years, SD = 11;2 years), recruited online through Prolific Academic (Prolific.ac). Several participants reported speaking more languages than just English[8]. For these languages, it was

[6] Three additional children were tested but were not included in the analysis: 2 did not complete the full experimental session due to technical failure, and 1 did not speak English as a native language.
For 2 children, a technical failure occurred in the identification task; since they completed the full training and part of the post-test, their data was used for analysis. Therefore, the analyses of the children's identification task are based on 46 participants, and all other analyses are based on 48 participants.

[7] One additional adult was tested, but did not complete the experimental session due to a technical error.

[8] 32 of the Online adults reported only speaking English, 10 additionally spoke French, 3 Spanish, 3 German, 2 Japanese, and 1 Korean, all at basic levels and mainly learnt through primary and secondary education. 5 of the Lab-based adults reported only speaking English, 27 additionally

checked whether they used the vowels that were part of training; this was not the case for any of the languages, and therefore, participants were not excluded. All participants had normal or corrected-to-normal vision, unimpaired hearing, and no dyslexia or language impairment[9]. Children were tested individually by a researcher in their school, and informed opt-in consent was obtained from a parent or caretaker prior to testing, while the adults recruited through the subject pool were tested individually in a sound-attenuated booth at UCL and signed a consent form before participating, and those recruited online were tested using the Gorilla experimental interface (www.gorilla.sc, Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed (2019)), through which informed consent was also obtained.

In order to avoid accidental differences in the starting ability between the two participant groups (as seen with adults in Experiment 1) participants were pseudo-randomly assigned to one of the 16 counterbalanced versions of the two experimental conditions on the basis of their score on the first training block (where training is identical for both conditions), with the goal of matching

spoke French, 14 German, 13 Spanish, 6 Mandarin, 3 Russian, 2 Bengali, 2 British Sign Language, 2 Punjabi, 1 Bangla, 1 Danish, 1 Hindu, 1 Japanese, 1 Korean, 1 Malay, 1 Swahili, 1 Urdu, and 1 Vietnamese. All children were learning French at school. 7 children reported speaking no other languages, 5 additionally spoke Jamaican, 5 Turkish, 4 Ghanaian, 4 Nigerian, 4 Tamil, 3 French, 2 Congolese, 2 an unspecified Ugandan language, 1 Bengali, 1 Czech, 1 Ethiopian, 1 Ewe, 1 Greek, 1 Igbo, 1 Lingala, 1 Lithuanian, 1 Malayalam, 1 Malian, 1 Mauritian, 1 Romanian, 1 Spanish, 1 Twi, 1 Urdu, and 1 Yoruba.

[9] One of the online adults reported having a delayed speech onset when younger.

performance in this test across conditions. Three participants were assigned to each of the 16 versions – see Table 4 above. For the adults who took part in the online replication, due to a slight glitch in the automated randomisation process 25 of the final set of participants were given the LV condition and 23 participants were given the HV condition. In return for their participation, children received stickers and a certificate, while adults received either 0.5 credits or a payment at a rate equivalent to £7.50 per hour: £4 for the adults tested in person, and £3.50 for the adults tested online.

Stimuli

Stimuli consisted of 12 monosyllabic two-way minimal pairs, all CVC-contexts containing the same Dutch vowels as in Experiment 1: <au> /ɔu/, <eu> /ø:/, and <ui> /œy/ (see Table 14). Since the minimal pair items used in Training needed to be imageable to make the task more like word-learning, two-way minimal pairs were chosen as there were not enough real-word minimal triplets available for these vowels. Six real-word minimal pairs were used in training, and 6 pseudo-word minimal pairs were used as novel items in the discrimination test (pre- and post). Consonant contexts were created according to the same constraints as in Experiment 1, and pseudo-word items were again matched to the real-word items in the overall place of articulation. The 12 real-word items presented in minimal pairs each had a corresponding clipart picture chosen from free online clipart

databases (see Appendix II). Stimuli recordings were made in the same sessions as for Experiment 1.

Vowel contrast	Trained real-word items	Novel pseudo-word items for Discrimination
AU-EU /ɔu/ - /ø:/	faun - fohn /fɔun/ - /fø:n/	fauf - feuf /fɔuf/ - /fø:f/
	kous - keus /kɔus/ - /kø:s/	maut - meut /mɔut/ - /mø:t/
EU-UI /ø:/ - /œy/	beuk - buik /bø:k/ - /bœyk/	keuf - kuif /kø:f/ - /kœyf/
	Teun - tuin /tø:n/ - /tœyn/	teup - tuip /tø:p/ - /tœyp/
AU-UI /ɔu/ - /œy/	koud - kuit /kɔut/ - /kœyt/	baus - buis /bɔus/ - /bœys/
	zout - zuid /zɔut/ - /zœyt/	sauk - suik /sɔuk/ - /sœyk/

Table 14. List of minimal pairs used in Experiment 2.

Design

The experimental design was nearly identical to that of Experiment 1, except that instead of mapping spoken words to shapes, this time each word used in training was accompanied by a clipart picture depicting its meaning. This meant that the shapes introduction task at pre-test was replaced by a vocabulary introduction task. Additionally, although this experiment continues to use *novel talkers* in both of the tests, it was impossible to use novel *items* in the Identification task as had been done for Experiment 1: if participants would be asked to map a novel word to either a picture they had seen before or a novel picture, this would not be testing their

vowel perception but rather their memory for the pictures used with the trained items.

Instead, in the Identification task an additional manipulation was included: minimal pairs (i.e. trials similar to training, where target and foil differ only in the key vowel) versus non-minimal pair test items (where the foil is not a minimal pair with the target). By including this task, it is possible to test participants' word learning more broadly; since Barcroft & Sommers (2005) and Sommers & Barcroft (2006) found broader word learning to also show a benefit of HV training input, this benefit should be expected here as well.

There were once more two variability conditions (HV and LV), and again the stimuli items used in all the tasks were identical across conditions. Speaker counterbalancing across the two variability conditions was identical to that used in Experiment 1 (see Table 4 above).

Procedure

Figure 13 shows the experimental trials for each of the tasks; the procedure for each task is described below. All tasks for the children and for the adults recruited through the UCL subject pool were run using PsychoPy (Peirce, 2007) on a laptop computer in quiet classrooms at the school or in a sound-attenuated booth at UCL. Adults participating in the online replication were tested through Gorilla (www.gorilla.sc , Anwyl-Irvine et al., 2019), with data being collected on 17 August 2018. Participants were able to run the experiment on a desktop or laptop

computer, or a tablet device, and were told to do the experiment in a quiet environment using a set of headphones. Stimuli were presented binaurally over headphones at a comfortable listening level.

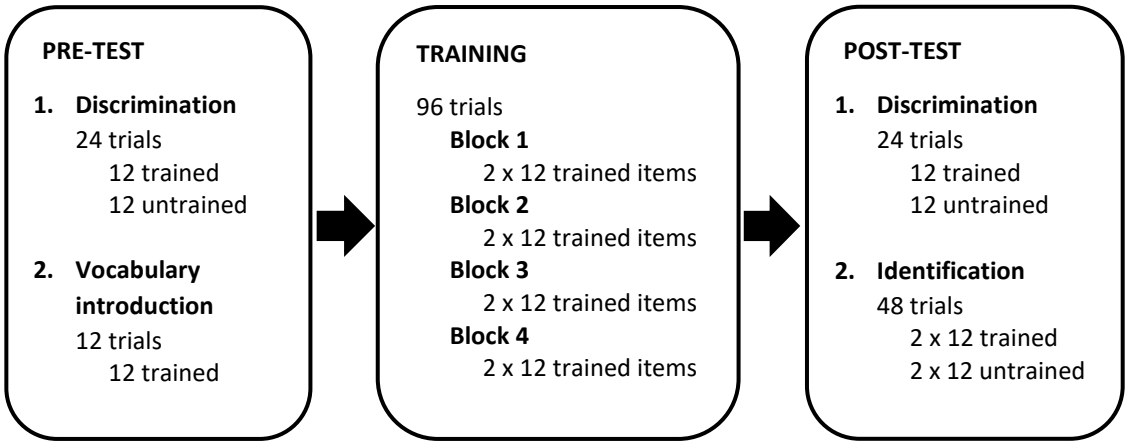


Figure 13. Number and type of experimental trials per task for experiment 2.

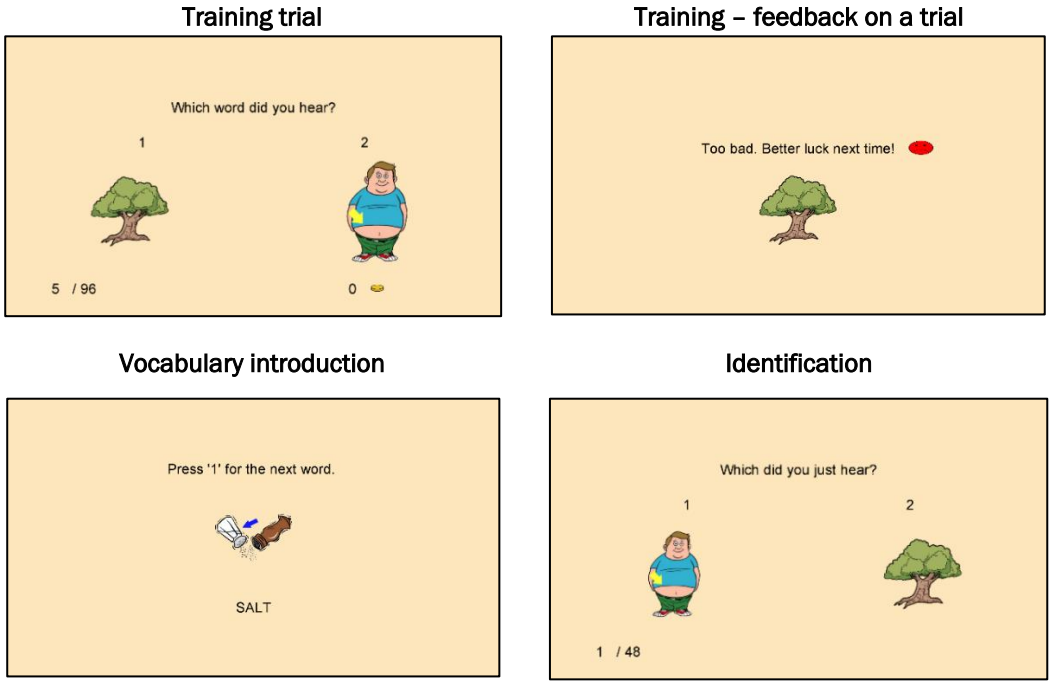


Figure 14. Stills from Training, Vocabulary introduction and the Identification task of Experiment 2.

Discrimination task

The procedure for the Discrimination task was identical to that of Experiment 1.

Vocabulary introduction

In the Vocabulary Introduction task, participants heard each of the 12 items they were to be trained on once in random order, and saw the corresponding clipart picture accompanied by an English translation of the Dutch word on the screen (see Figure 14). This task was included to familiarise participants with the words and pictures that would occur in training, and to ensure they understood the meaning of each picture.

Training

The training procedure was identical to that of Experiment 1, except that rather than shapes matching the stimuli, the clipart pictures that matched the stimuli were displayed on the screen. Additionally, because two-way rather than three-way minimal pair contrasts were used as stimuli, the training task was a 2-alternative forced choice task, rather than a 3-alternative forced choice task as in Experiment 1 (see Figure 14).

Identification task

The procedure of this task was identical to the training task except that (a) no feedback was given (and no coins were received), and (b) half the trials were presented in the trained minimal pairs while half were presented in untrained combinations; this distinction was made to be able to test whether participants had learnt the words, as discussed above. Presentation of minimal pairs and non-minimal pairs was blocked, with trials presented in a randomised order within each

section. All stimuli were spoken by a novel talker and were repeated once. See Figure 14 for a scene from the identification task.

Analyses

The same approach was used as in Experiment 1 and the same hypotheses were tested as outlined in Table 5 except that, in the Identification task, the two hypotheses concerning novel items could not be tested as the task did not contrast novel versus trained items. The approach for this test was to test minimal pair and non-minimal pair test items separately, and to test the same two hypotheses for each (i.e. that performance is above chance, and that performance is greater with HV performance). Again analyses were conducted separately for children and adults, and here the data from the online and lab-based participants were also analysed separately. In addition, where differences between groups were tested, comparisons of the online and lab-based participants were included. As before, for completeness sake, all effects related to the hypotheses described above will be reported, but not all will be interpreted.

3.3.2 Results

Adult data

Discrimination task

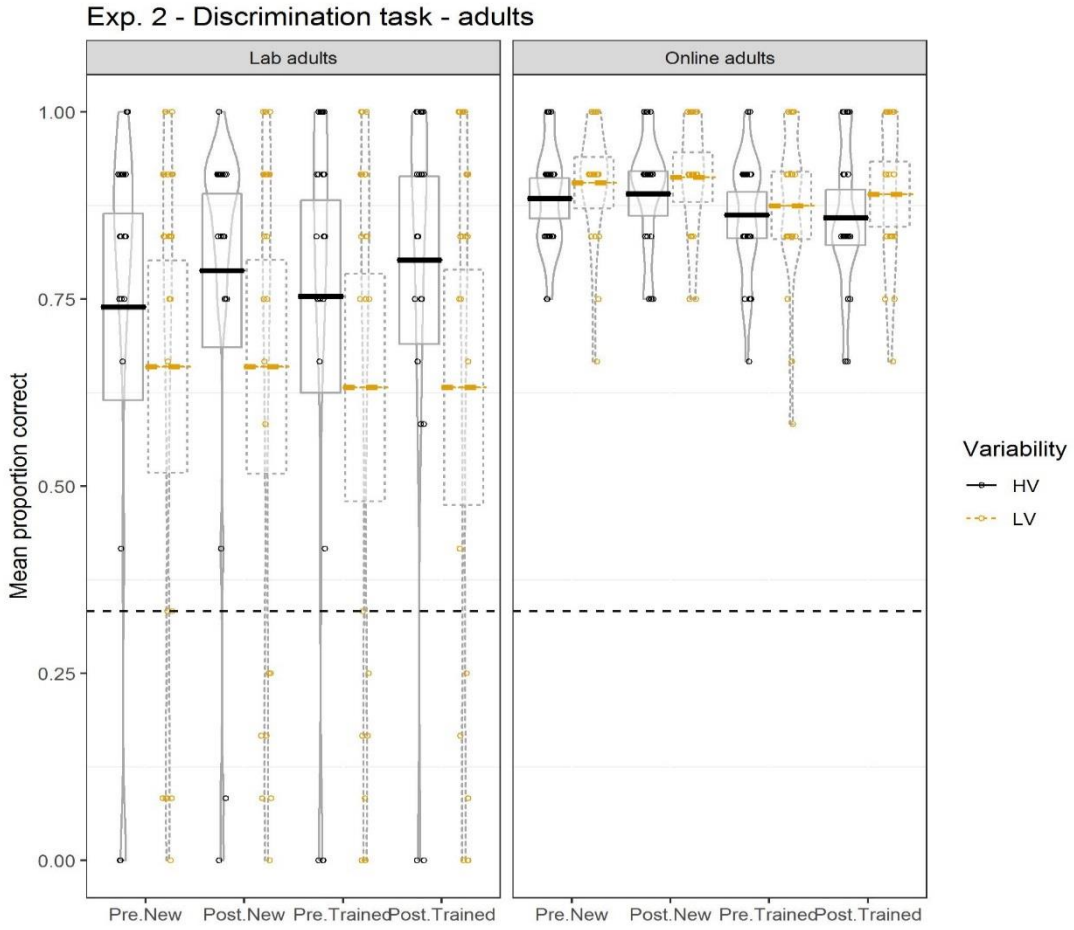


Figure 15. Accuracy results for Lab-based adults (on the left) and Online adults (on the right) on the pre- and post-test discrimination task of Experiment 2, comparing accuracy for new versus trained items as well as HV versus LV. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

The data for lab-based and online adult participants are plotted in Figure 15. Statistics for the key hypotheses tested are shown in Table 15 and Table 16. For both groups, it can be seen that while there is strong evidence that participants are above chance on this test (lab-based 71%, online 89% compared with 33% chance), there is substantial evidence that they do *not* improve from pre- to post-training

(lab-based: 70% to 72%; online: 88% to 86%). There is ambiguous evidence for there being a difference in improvement for trained and untrained items for lab-based adults, while there is substantial evidence against such a difference for online adults. Turning to the key hypotheses concerning variability, lab-based adults show ambiguous evidence for greater improvement in the HV condition, and substantial evidence against HV being particularly beneficial for novel items, while online adults show substantial evidence against greater improvement in HV, and ambiguous evidence for an HV benefit for novel items.

Lab-based participants

Final structure Exp2_LabAdult_DM model: accuracy ~ session * condition * itemnovelty + speaker + (session*itemnovelty|participant) + (session|contrast)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	2.098	0.425	4.934	<.001	0.475	1145.861	[0.1823, >3.8223]
improvement from pre- to post- test	session	0.126	0.214	0.591	.554	1.539	0.233	[1.0823, ∞]
greater improvement for trained items	item-novelty : session	0.189	0.318	0.595	.552	1.539	0.745	[0, 3.6423]
greater overall improvement for HV training	condition: session	0.430	0.388	1.107	.268	1.539	0.343	[0, 1.5823]
greater improvement in HV greater for untrained items	*condition : item-novelty : session	-0.050	0.557	-0.090	.928	1.539	0.316	[1.4623, ∞]

Table 15. Mixed model results for the Discrimination analysis of Experiment 2, for Lab-based adults. *The sign of these fixed effects is changed from the model coding to be in line with the hypothesis.

Online participants

Final structure Exp2_OnlineAdult_DM model: accuracy ~ session * condition * itemnovelty + speaker + (session*itemnovelty|participant) + (session|contrast)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	3.316	0.265	12.530	<.001	0.475	$2.57*10^{225}$	[0.1823, >3.8223]
improvement from pre- to post- test	session	0.090	0.193	0.466	.642	1.539	0.185	[0.8423, ∞]
greater improvement for trained items	item-novelty : session	-0.021	0.306	-0.068	.946	1.539	0.183	[0.8323, ∞]
greater overall improvement for HV training	condition: session	-0.124	0.291	-0.425	.671	1.539	0.135	[0.6023, ∞]
greater improvement in HV greater for untrained items	*condition : item-novelty : session	0.181	0.581	0.312	.755	1.539	0.451	[0, 2.1623]

Table 16. Mixed model results for the Discrimination analysis of Experiment 2, for Online adults. *The sign of these fixed effects is changed from the model coding to be in line with the hypothesis.

Training

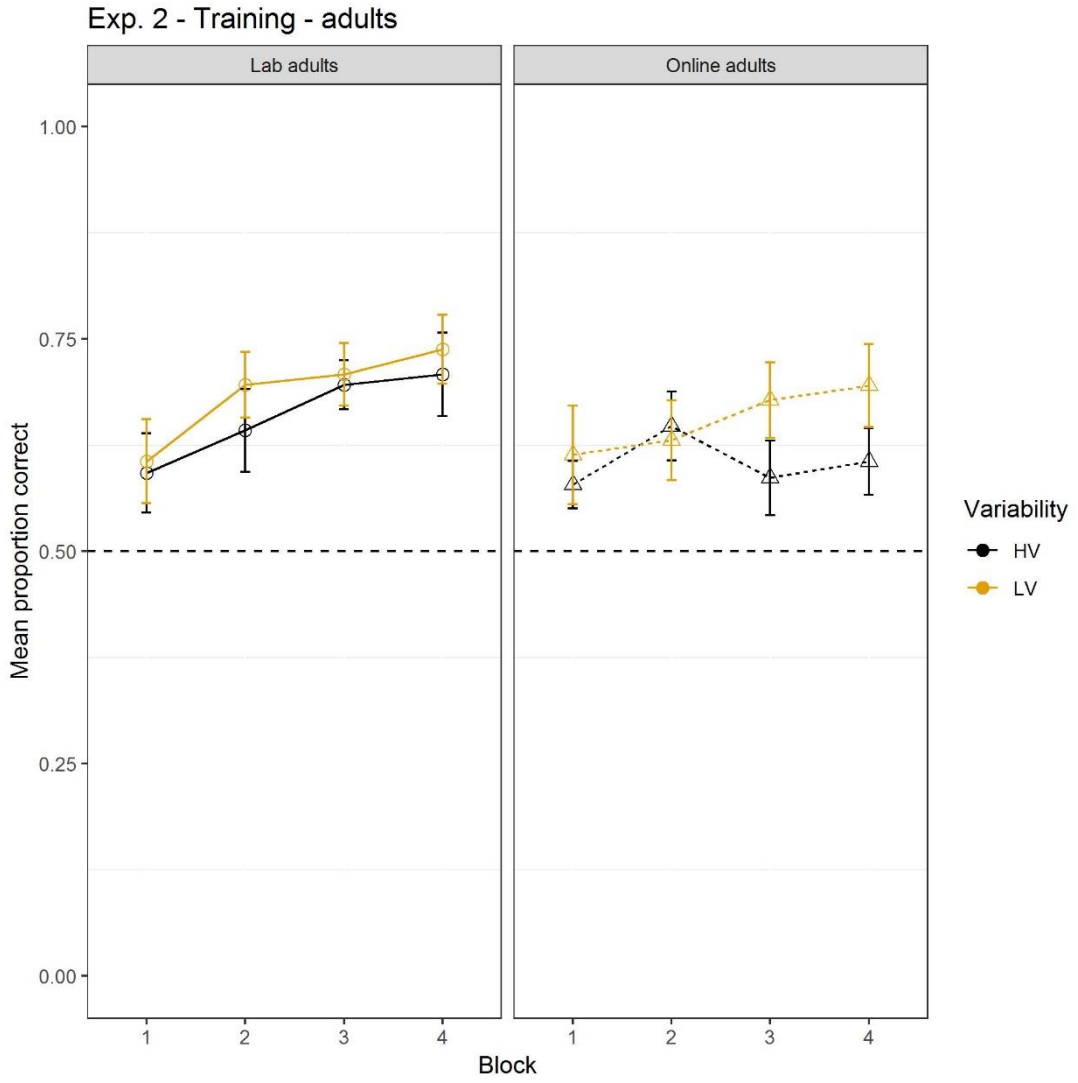


Figure 16. Accuracy results for Lab-based adults (on the left) and Online adults (on the right) in training of Experiment 2 plotted over the 4 blocks, comparing HV and LV. Error bars show 95% CI, and the dashed line indicates chance level.

The data for lab-based and online adult participants are plotted in Figure 16. Statistics for the key hypotheses tested are shown in Table 17 and Table 18. For both groups, there is substantial evidence that adults are above chance, and that they improve across the blocks. Online adults show substantial evidence for greater improvement in the LV condition, while for lab-based adults this evidence is ambiguous.

Lab-based participants

Final structure Exp2_LabAdults_Train model: accuracy ~ block * condition + speaker + (block|participant) + (block |word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.795	0.138	5.759	<.001	0.544	$2.87*10^6$	[0.0417, >4.5517]
improvement across blocks	block	0.211	0.043	4.889	<.001	0.07	6075.806	[0.0417, >4.5517]
greater improvement in LV condition	*condition: block	0.012	0.069	0.168	0.867	0.131	0.526	[0.0417, 0.2217]

Table 17. Mixed model results for the Training analysis of Experiment 2, for Lab adults. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Online participants

Final structure Exp2_OnlineAdults_Train model: accuracy ~ block * condition + speaker + (block|participant) + (block|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.544	0.088	6.182	<.001	0.795	$3.5*10^7$	[0.0417, >4.5517]
improvement across blocks	block	0.070	0.031	2.303	.021	0.211	3.813	[0.0417, 0.2417]
greater improvement in LV condition	*condition: block	0.131	0.060	2.177	.029	0.07	4.852	[0.0417,0.3917]

Table 18. Mixed model results for the Training analysis of Experiment 2, for Online adults. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Figure 17 below shows the degree of variability in the patterns of change over blocks across lab-based adults. As can be seen, the majority of lab-based adults

show a similar pattern of learning across blocks, though the steepness of learning differs across participants.

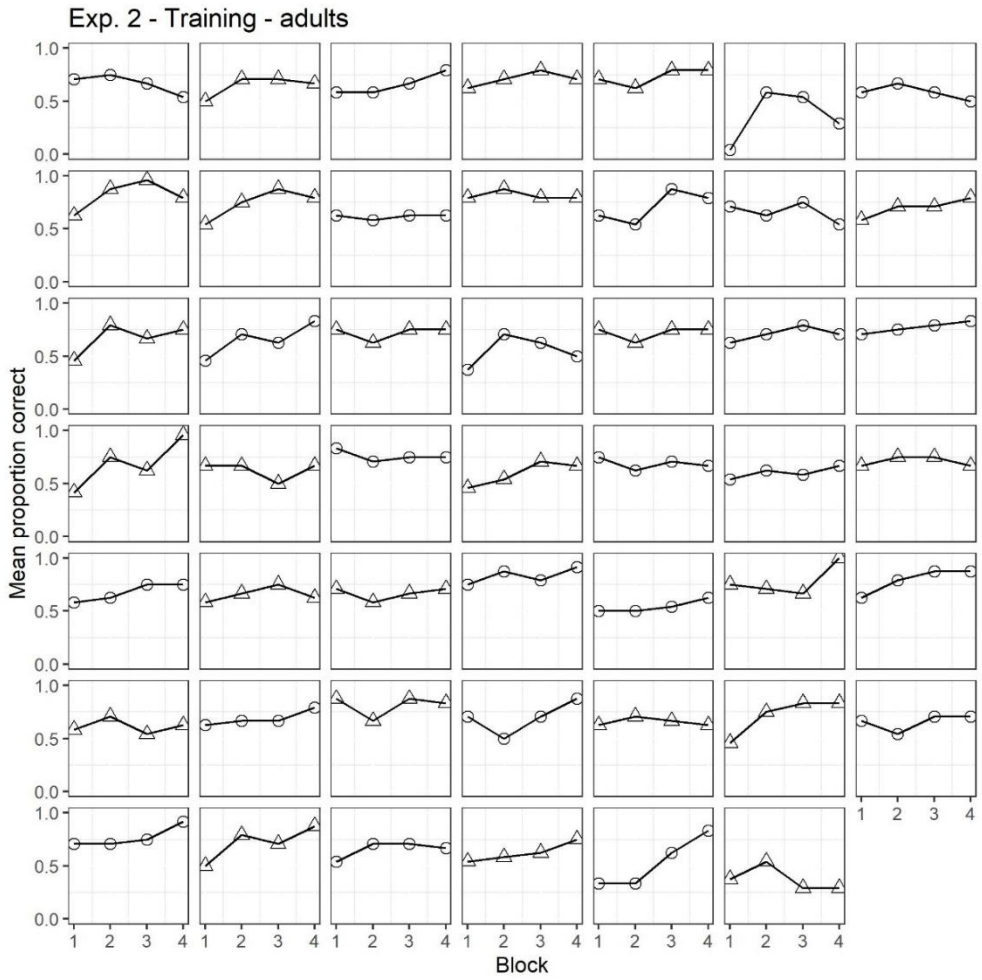


Figure 17. Lab-based adult accuracy scores in training of Experiment 2 plotted over the 4 blocks, split out by participant. Plots with circles indicate a participant received HV training, while triangles indicate LV training.

Figure 18 shows the pattern of improvement across the training blocks for lab-based adults split out by vowel: while performance on <ui> is slightly lower than that for <au> and <eu>, the pattern of improvement is similar.

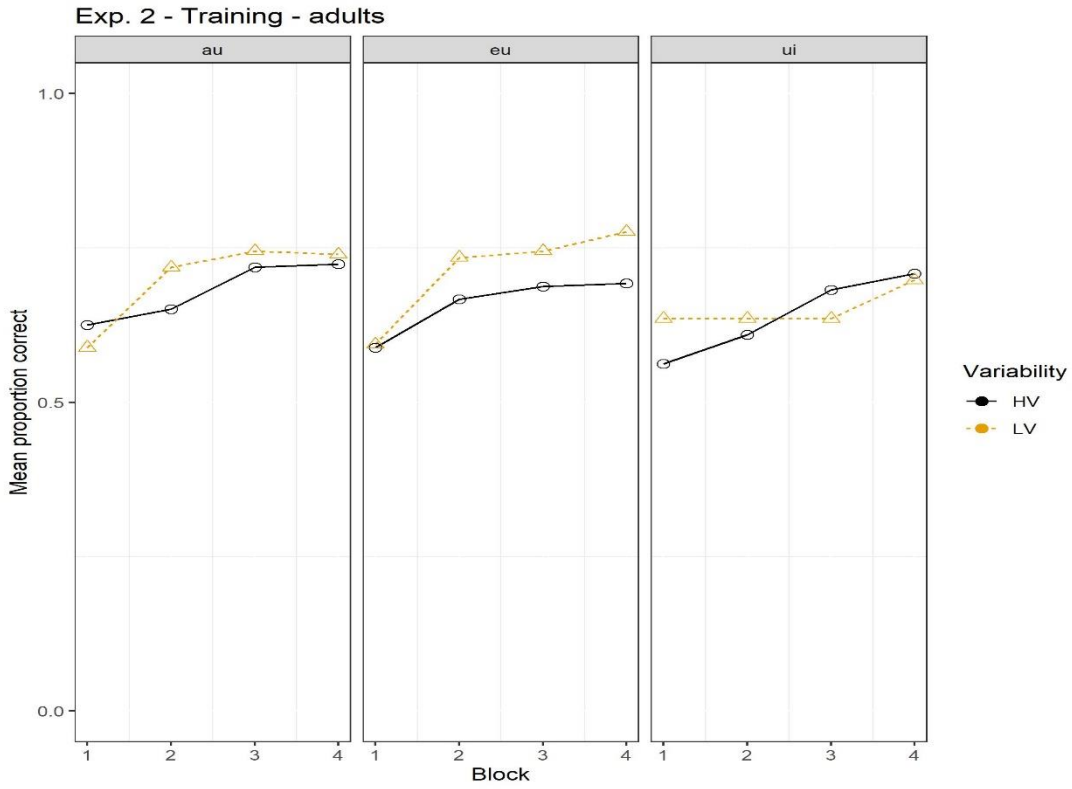


Figure 18. Lab-based adult accuracy scores in training of Experiment 2 plotted over the 4 blocks, split out by vowel, comparing HV and LV input.

Identification task (post-test only)

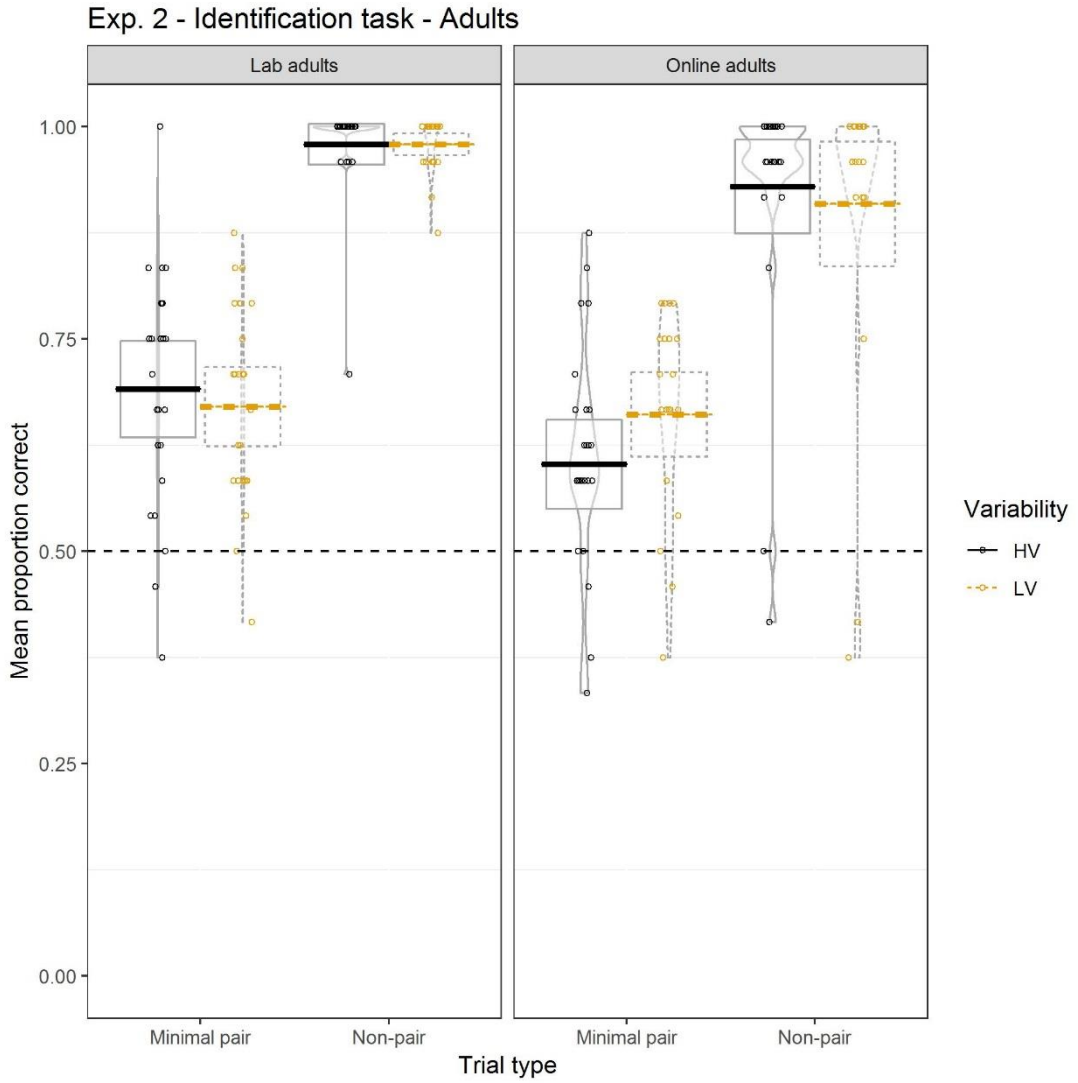


Figure 19. Accuracy results for Lab-based adults (on the left) and Online adults (on the right) on the identification task of Experiment 2, comparing items presented in minimal pairs and non-pairs in the two variability conditions. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 19 and Table 19, Table 20, Table 21, and Table 22, for both groups there was substantial evidence for adults being above chance in the minimal pairs as well as the non-minimal pairs. Lab-based participants show ambiguous evidence for HV outperforming LV for both minimal pairs and non-minimal pairs, while

online participants show substantial evidence against HV outperforming LV in both the minimal and the non-minimal pairs.

Lab-based participants

Minimal pairs

Final structure Exp2_LabAdult_ID_minpair model: accuracy ~ condition + speaker + (condition|participant) + (1|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.851	0.192	4.436	<.001	0.584	4481.144	[0.0834, >3.8434]
greater performance in HV condition	condition	0.137	0.189	0.722	.471	0.851	0.421	[0, 1.1034]

Table 19. Mixed model results for the minimal pair Identification analysis of Experiment 2, for Lab-based adults.

Non minimal pairs

Final structure Exp2_LabAdult_ID_nonpair model: accuracy ~ condition + speaker + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	6.146	1.255	4.898	<.001	3.661	29823.551	[0.2834, >3.8434]
greater performance in HV condition	condition	3.603	2.262	1.593	.111	6.146	1.967	[0, >3.8434]

Table 20. Mixed model results for the non-minimal pair Identification analysis of Experiment 2, for Lab-based adults.

Online adults

Minimal pairs

Final structure Exp2_OnlineAdult_ID_minpair model: accuracy ~ condition + speaker + (1|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.584	0.174	3.363	.001	0.851	91.098	[0.0834, >3.8434]
greater performance in HV condition	condition	-0.234	0.191	-1.228	.220	0.584	0.151	[0.2334, ∞]

Table 21. Mixed model results for the minimal pair Identification analysis of Experiment 2, for Online adults.

Non minimal pairs

Final structure Exp2_OnlineAdult_ID_nonpair model: accuracy ~ condition + speaker + (condition|participant) + (1|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	3.661	0.375	9.763	<.001	6.146	$5.08*10^{19}$	[0.0834, >3.8434]
greater performance in HV condition	condition	0.101	0.752	0.134	.893	3.661	0.222	[2.3834, ∞]

Table 22. Mixed model results for the non-minimal pair Identification analysis of Experiment 2, for Online adults.

Child data

Discrimination task

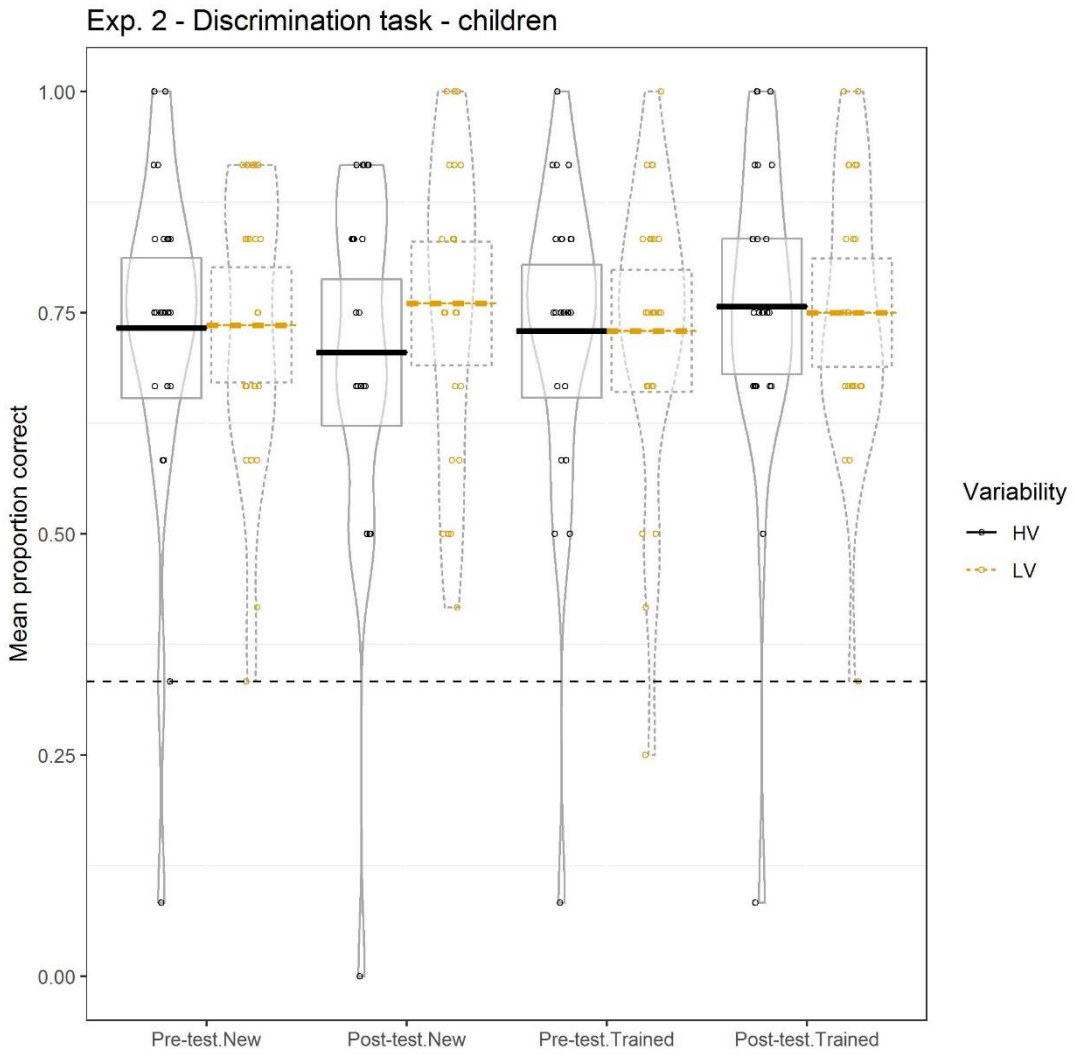


Figure 20. Accuracy results for children on the pre- and post-test discrimination task of Experiment 2, comparing accuracy for new versus trained items as well as HV versus LV. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

There was substantial evidence that children are above chance in Discrimination, as shown in Figure 20 and Table 23. There was substantial evidence against an improvement from pre- to post-test, and against more improvement in HV than LV as well as vice versa. There was substantial evidence against more improvement

in trained than untrained items, and against more improvement in HV for untrained items.

Final structure Exp2_Child_DM model: accuracy ~ session * condition * itemnovelty + speaker + (session*itemnovelty|participant) + (session|contrast)

Hypothesis	fixed effect in model	beta	SE	Z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	1.983	0.226	8.779	<.001	2.029	$4.15*10^{15}$	[0.1823, >3.8223]
improvement from pre- to post- test	session	0.087	0.117	0.739	.460	1.665	0.14	[0.7023, ∞]
greater improvement for trained items	item-novelty : session	0.127	0.225	0.564	.573	1.665	0.221	[1.1023, ∞]
greater overall improvement for HV training	condition: session	-0.162	0.218	-0.744	.457	1.665	0.077	[0.3623, ∞]
greater overall improvement for LV training	*condition: session	0.162	0.218	0.744	.457	1.665	0.261	[1.2923, ∞]
greater improvement in HV for untrained items	*condition : item-novelty : session	-0.440	0.440	-1.000	.317	1.665	0.134	[0.6323, ∞]

Table 23. Mixed model results for the Discrimination analysis of Experiment 2, for Children. *The sign of these fixed effects is changed from the model coding to be in line with the hypothesis.

Training

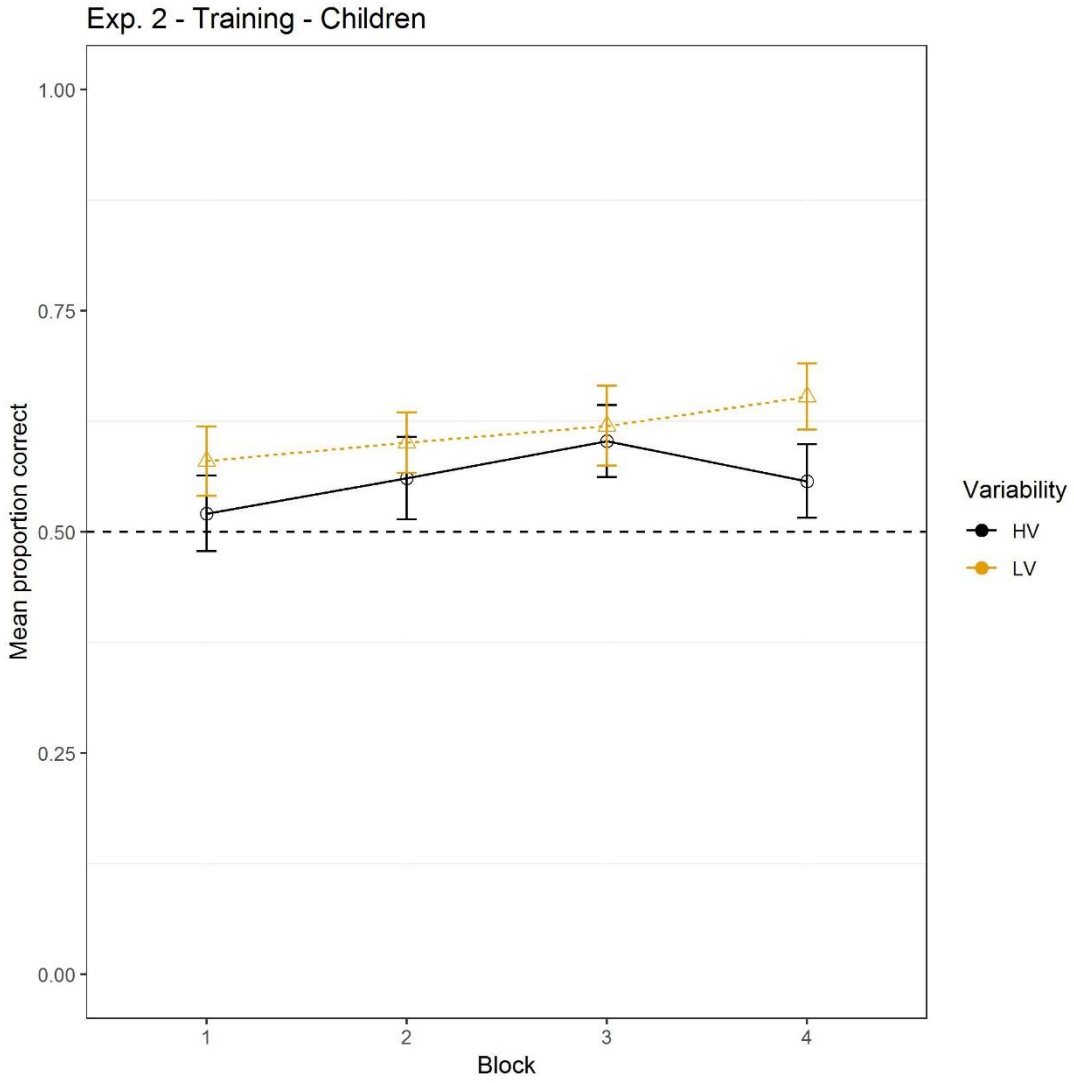


Figure 21. Child accuracy scores in training of Experiment 2 plotted over the 4 blocks, comparing HV and LV. Error bars show 95% CI, and the dashed line indicates chance level.

There was substantial evidence for children being above chance in training, and for improvement across blocks, indicating there was an effect of training (see Figure 21 and Table 24). The evidence for greater improvement in LV than HV was ambiguous.

Final structure Exp2_Child_Train model: accuracy ~ block * condition + speaker + (block|participant) + (condition*block |word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.359	0.049	7.388	<.001	0.544	$1.02*10^{11}$	[0.0417, >4.5517]
improvement across blocks	block	0.083	0.039	2.143	.032	0.07	5.448	[0.0417, 0.2217]
greater improvement in LV condition	*condition: block	0.037	0.056	0.661	.509	0.131	0.686	[0, 2917]

Table 24. Mixed model results for the Training analysis of Experiment 2, for Children. *The sign of these fixed effects is changed from the model coding to be in line with the hypothesis.

Figure 22 below shows the degree of variability in the patterns of change over blocks across participants. As can be seen, the majority of children do show a similar pattern of learning across blocks for the picture-based paradigm this time.

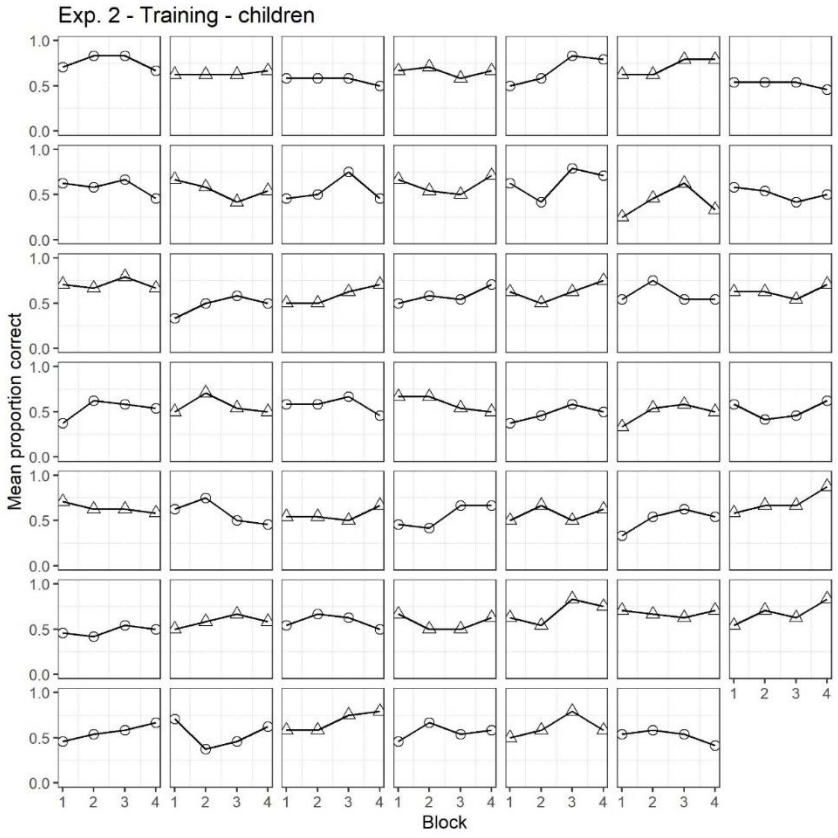


Figure 22. Child accuracy scores in training of Experiment 2 plotted over the 4 blocks, split out by participant. Plots with circles indicate a participant received HV training, while triangles indicate LV training.

Figure 23 shows the pattern of improvement across the training blocks split out by vowel: while performance on <ui> is slightly lower and less steep than that for <au> and <eu>, the pattern of improvement is similar.

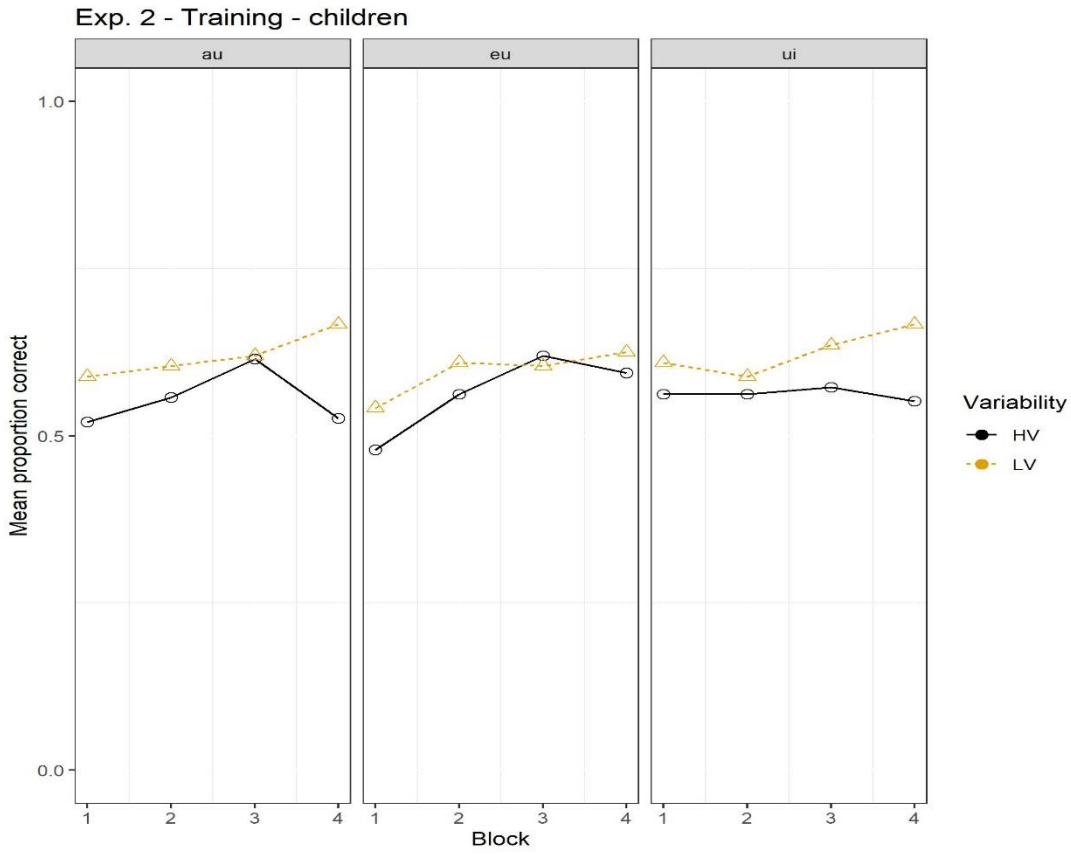


Figure 23. Child accuracy scores in training of Experiment 2 plotted over the 4 blocks, split out by vowel, comparing HV and LV input.

Identification task (post-test only)

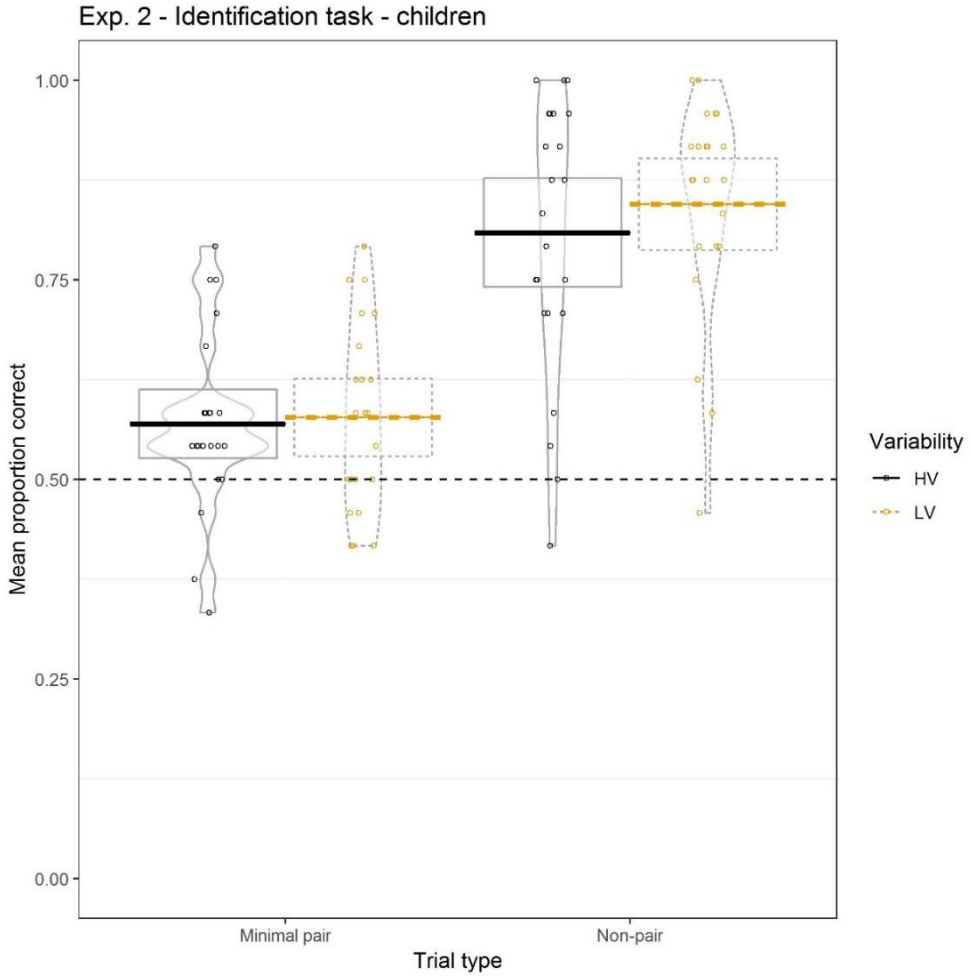


Figure 24. Child accuracy results on the identification task of Experiment 2, comparing items presented in minimal pairs and non-pairs in the two variability conditions. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 24, Table 25 and Table 26, there is substantial evidence that children are above chance in minimal pairs as well as non-pairs. Turning to the key hypotheses concerning variability, there was substantial evidence against greater performance in HV over LV for minimal pairs while this was ambiguous in non-pairs, and ambiguous evidence for greater performance in LV over HV in both minimal pairs and non-pairs.

Minimal pairs

Final structure Exp2_Child_ID_minpair model: accuracy ~ condition + speaker + (1|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance	intercept-chance	0.298	0.066	4.531	<.001	0.584	102.837	[0.0834, >3.8434]
greater performance in HV	condition	-0.034	0.137	-0.248	.804	0.298	0.275	[0.2434, ∞]
greater performance in LV	Condition	0.034	0.137	0.248	.804	0.298	1.561	[0, 2.1434]

Table 25. Mixed model results for the minimal pair Identification analysis of Experiment 2, for children.

Non minimal pairs

Final structure Exp2_Child_ID_nonpair model: accuracy ~ condition + speaker + (condition|participant) + (1|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance	intercept-chance	1.899	0.204	9.308	<.001	3.661	$6.18*10^{19}$	[0.0834, >3.8434]
greater performance in HV	condition	-0.146	0.356	-0.409	.682	1.899	0.41	[0, 2.3734]
greater performance in LV	condition	0.146	0.356	0.409	.682	1.899	0.336	[0, 1.8934]

Table 26. Mixed model results for the minimal pair Identification analysis of Experiment 2, for children.

Comparing different groups

The analyses reported above suggest some differences between lab-based adults, online adults, and children. As before, wherever an effect was found in at least one age group, it is of interest to compare the participant groups. In addition to comparing both adult groups to the children's performance, it is methodologically interesting to see if the lab-based and online adults differ. The hypothesis here is

that lab-based adults outperform online adults. Results are in Table 27, Table 28, and Table 29.

For child versus adult comparisons: there was substantial evidence that lab-based adults did not outperform children on Discrimination, while online adults did. In Training, both lab-based and online adults showed substantial evidence for outperforming children. Lab-based adults also showed greater improvement across blocks than children, while for online adults there was evidence for the null. Although there was evidence that the online adult group improved more in training with LV items than HV items, there was no evidence that this effect differed from that in children (where it was ambiguous). Both groups of adults substantially outperformed children in the Identification task, for both minimal pair trials and non-pair trials.

For the comparison of the two adults groups: this showed evidence against lab-based adults outperforming online adults in Discrimination (note that the significant *p*-value in Table 29 reflects an effect in the *reverse* direction). Despite this, there was substantial evidence for lab-based adults outperforming online adults on training overall, and improving more across blocks. Although the separate analyses only found substantial evidence for an LV benefit in Training for lab-based adults, with the evidence ambiguous for online adults, the evidence for the comparison between groups was ambiguous. Lab-based adults outperformed online adults on the Identification task for the non-pair trials, but evidence was ambiguous for the minimal pair trials.

Lab adults versus children

Final structure Exp2_LabAdultvsChild_DM model: accuracy ~ session * condition * itemnovelty + group + speaker + (session*itemnovelty|participant) + (session|contrast)

Final structure Exp2_LabAdultvsChild_Train model: accuracy ~ block * condition * group + speaker + (block|participant) + (condition*block|word)

Final structure Exp2_LabAdultvsChild_ID_minpair model: accuracy ~ condition + speaker + group + (1|participant) + (1|word)

Final structure Exp2_LabAdultvsChild_ID_nonpair model: accuracy ~ condition + speaker + group + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
lab adults outperform children in discrimination	Discrimination model: *group*	-0.187	0.081	-2.313	.021	0.341	0.069	[0, ∞]
lab adults outperform children in training	Training model: *group*	0.387	0.045	8.658	<.001	0.086	$6.02*10^{12}$	[0, >4.5517]
lab adults show greater increase in performance across blocks than children in training	Training model: *group by block*	0.108	0.040	2.718	.007	0.186	14.237	[0, 1.0017]
lab adults show greater performance than children in identification task for minimal pairs	Identification minpair model: *group*	0.469	0.088	5.300	<.001	0.077	640.704	[0.0834, >3.8434]
lab adults show greater performance than children in identification task for non-minimal pairs	Identification nonpair model: *group*	2.627	0.296	8.883	<.001	0.267	$5.74*10^{7}$	[0.0834, >3.8434]

Table 27. Mixed model results for the Lab adults versus Children Comparison of Experiment 2.

Online adults versus children

Final structure Exp2_OnlineAdultvsChild_DM model: accuracy ~ session * condition * itemnovelty + group + speaker + (session*itemnovelty|participant) + (session|contrast)

Final structure Exp2_OnlineAdultvsChild_Train model: accuracy ~ block * condition * group + speaker + (block|participant) + (condition*block|word)

Final structure Exp2_OnlineAdultvsChild_ID_minpair model: accuracy ~ condition + speaker + group + (condition|participant) + (condition|word)

Final structure Exp2_OnlineAdultvsChild_ID_nonpair model: accuracy ~ condition + speaker + group + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
online adults outperform children in discrimination	Discrimination model: *group*	1.203	0.176	6.835	<.001	0.341	$9.34*10^7$	[0.1823, >3.8223]
online adults outperform children in training	Training model: *group*	0.186	0.068	2.737	.006	0.086	12.199	[0, 1.8917]
online adults show greater increase in performance across blocks than children in training	Training model: *group by block*	-0.012	0.041	-0.283	.778	0.108	0.293	[0.1017, ∞]
online adults show greater LV benefit for increase in performance across blocks than children in training	Training model: *group by block by variability*	-0.087	0.082	-1.053	.292	0.029	1.003	[0, 0.7317]
online adults show greater performance than children in identification task for minimal pairs	Identification minpair model: *group*	0.251	0.108	2.327	.020	0.077	3.698	[0.0834, 1.0234]
online adults show greater performance than children in identification task for non-minimal pairs	Identification nonpair model: *group*	1.441	0.354	4.066	<.001	0.267	31.676	[0.1034, >3.8434]

Table 28. Mixed model results for the Online adults versus Children Comparison of Experiment 2.

Lab adults versus online adults

Final structure Exp2_LabAdultvsOnlineAdult_DM model: accuracy ~ session * condition * itemnovelty + group + speaker + (session*itemnovelty|participant) + (session|contrast)

Final structure Exp2_LabAdultvsOnlineAdult_Train model: accuracy ~ block * condition * group + speaker + (block|participant) + (condition*block|word)

Final structure Exp2_LabAdultvsOnlineAdult_ID_minpair model: accuracy ~ condition + speaker + group + (1|participant) + (1|word)

Final structure Exp2_LabAdultvsOnlineAdult_ID_nonpair model: accuracy ~ condition + speaker + group + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
lab adult outperform online adults in discrimination	Discrimination model: *group*	-1.374	0.352	-3.902	<.001	0.341	0.188	[0, ∞]
lab adult outperform online adults in training	Training model: *group*	0.214	0.086	2.492	.013	0.086	6.42	[0.0517, 1.2317]
lab adult show greater increase in performance across blocks than online adults in training	Training model: *group by block*	0.127	0.046	2.771	.006	0.223	15.987	[0, 1.3717]
lab adults show greater LV benefit for increase in performance across blocks than online adults in training	Training model: *group by block by variability*	0.122	0.092	1.327	.185	0.029	0.713	[0, 0.1017]
lab adults show greater performance than online adults in identity task minimal pairs	Identification model: *group*	0.245	0.126	1.954	.051	0.077	2.433	[0, 0.1034]
lab adults show greater performance than online adults in identity task non-pairs	Identification model: *group*	1.465	0.488	3.001	.003	0.267	4.579	[0.2034, >3.8434]

Table 29. Mixed model results for the Lab adults versus Online Adults Comparison of Experiment 2.

3.3.3 Discussion

In this second experiment, native English speaking 7-8 year-olds and adults recruited in the lab and online were trained, through means of a phonetic training paradigm, to associate meaningful pictures to the vowels in Dutch minimal triplets. Input variability was still the key manipulation, and performance was again measured through means of a discrimination and identification task. The hope was to see greater evidence of learning with this new paradigm, particularly for children who showed very little learning in Experiment 1. If evidence of learning could be seen, this might also allow for differences between variability conditions to emerge, in terms of a potential HV benefit for generalisation. In fact, significant evidence of learning and generalisation was found in all but the Discrimination test, but no evidence of a variability difference was found. Note that although the evidence of learning was significant, percentual improvement was still rather weak (around 5%) compared to that found in adults (between 10 and 20%, in line with the usual amounts of improvement found in phonetic training studies). The remainder of this discussion will briefly consider how the findings in each task differed from those found in Experiment 1, before considering the implications of the findings.

In Training, performance for both adult groups and children was above chance, and all groups showed improvement over training. This is in contrast to Experiment 1, where there was substantial evidence that children did not improve over training. Training results were further investigated to see if there was a

steeper training slope in the LV condition than the HV condition, due to adapting to a single speaker. There was only substantial evidence for this in the online adult participants, while the evidence was ambiguous for the lab-based adults and children.

Turning to the tests of generalization, the results of the discrimination test were similar to the previous experiment i.e. although both adult groups and children performed above chance, as in Experiment 1, no group showed improvement after training (with substantial evidence for the null in all groups); it is thus not sensible to interpret any further effects. The discussion of why training does not lead to improvement in Discrimination will be left for the general discussion in Section 3.5. In the second generalisation task, the Identification post-test which involved generalisation to new speakers, all groups were above chance, contrasting to Experiment 1 where children's performance was at chance, with substantial evidence for the null. This was the case for both minimal pairs and non-minimal pairs. Note that performance on the minimal pairs was still relatively weak, but clear enough to expect further learning to occur if the paradigm were to be used in a full-length training study. Despite finding above chance performance in this experiment, no group showed a HV benefit, with substantial evidence for the null in non-pairs for online adults and minimal pairs for children, and ambiguous evidence in all other cases. It is worth noting that in this experiment – unlike in Experiment 1 – there were no accidental differences between the age groups apparent at block 1, so the lack of an HV advantage can this time not be explained

by an accidental LV benefit preventing its appearance. For children, the alternative hypothesis that there was a benefit of LV training was also tested, with ambiguous evidence for both minimal pairs and non-pairs.

Overall, it seems that native English adults and children can learn to associate words containing novel Dutch vowels with the pictures, generalising across voices at test. Thus this study met the aim of promoting stronger one-session learning in children. Importantly though, there was still no evidence that high variability training promoted generalisation to novel voices and items, with substantial evidence against this hypothesis in children. In light of this key question concerning the role of variability, the fact that there is no evidence for an effect of variability on generalisation is particularly unexpected and contradicts some of the literature reviewed in Section 1.3.5 above. This key finding will be reviewed further in the general discussion in Section 3.5.

Another set of findings in this study relates to the differences between the different participant groups. Age differences are of theoretical interest given the literature, and in addition, for methodological reasons, it is also interesting to see whether there are differences in performance between adults tested in the lab, and adults recruited through online measures. Generally speaking, both adult groups outperformed children, though whether the evidence met the criteria for 'substantial' rather than ambiguous evidence differed for the different tests. This included evidence that adults outperformed children in terms of the steepness of their learning slopes in training (with substantial evidence for this comparison

with lab based adults), suggesting not just baseline ability but also learning speed is improved with age. Comparison of performance differences between adults tested in the lab and online showed that lab-tested adults substantially outperformed online adults on all but the Discrimination task. Thus there is at least some evidence that lab-based participants do outperform online participants, although note that fairly consistent patterns of performance are seen across the two groups. The fact that lab-based participants did not outperform online performance in Discrimination (with evidence for the null) is unexpected. Inspecting the graphs, it seems to be caused by more variable performance in lab-based adults, with online adults actually performing better on this task. This is hard to interpret; however, note that this data suggests that the idea that lab-based participants are better than online participants is not always true.

A key finding in this study is that, unlike in Experiment 1, child participants showed generalisation to novel speakers and items in learning to associate the non-native vowel contrasts. This is likely due to the change to a task in which participants associate pictures to *whole words*, rather than associating symbols to individual vowels. This result is somewhat surprising and potentially interesting since participants are actually having to learn *more* associations in this experiment: in learning to associate the vowels to symbols, they are learning 3 associations, while learning to map to individual pictures means they are learning 12. However, before conclusions can be drawn about this, first there is an unintended difference between Experiment 1 and Experiment 2 that needs to be resolved.

This unintended procedural difference occurs in the training task, and as a result also in the identification task in pre/post-test: Experiment 2 used a two-alternative forced choice task (2AFC), while Experiment 1 used a three-alternative forced choice task (3AFC). This difference came about due to restrictions on the stimuli: in Experiment 2 there were not sufficient real-word three-way minimal triplets that were both depictable and child-friendly. As this study aimed to teach participants real Dutch (as discussed above), creating novel Dutch pseudowords was not an option. Therefore, the choice was made to move to using two-way minimal pairs as stimuli instead, changing the training task and identification task from a 3AFC to a 2AFC. However, this could potentially have lowered task demands. The purpose of Experiment 3 was to establish whether the use of meaningful pictures is better than abstract symbols, even when task demands are controlled for, by repeating Experiment 1 with a 2AFC task.

3.4 Experiment 3

Experiment 3 used the 2AFC task used in Experiment 2, but with the symbol stimuli from Experiment 1. If the change between Experiment 1 and 2 was driven by the 2AFC task being easier than the 3AFC task, this should show up clearly in the child data as evidence of learning, indicated by improvement over Training and above chance performance in the Identification task. There should also be a substantial evidence that performance in the experiments differs in this respect. If

the change was due to the use of pictures, no such evidence of learning should be found.

3.4.1 Method

Participants

Participants were 48 native English speaking children[10] (mean age = 8;4 years, SD = 4 months) recruited from two primary schools in North London, and 48 adult native speakers of English[11] located in the UK or Ireland (mean age = 33;11 years, SD = 11;4 years) recruited through Prolific Academic (Prolific.ac). Several participants reported speaking more languages than just English[12]. For these languages, it was checked whether they used the vowels that were part of training; this was not the case for any of the languages, and therefore, participants were not excluded. All participants had normal or corrected-to-normal vision, unimpaired hearing, and none were dyslexic or had a language impairment[13]. Children were

[10] Nine additional children were tested but their data were not included in the analysis: 4 did not complete the full experimental session due to technical failure, 4 had a diagnosed language impairment, and 1 did not speak English as a native language.

[11] Three additional adult participants were tested, but did not complete the experimental session due to technical difficulties.

[12] 26 participants reported only speaking English; 11 participants reported speaking French, 4 German, 4 Spanish, 2 Irish, and 1 Scottish Gaelic, all at basic levels, and mainly learnt through secondary education. All children were learning French at school. 8 children reported speaking no other languages, 10 additionally spoke Greek, 6 Bengali, 5 Arabic, 4 Turkish, 4 Spanish, 2 Italian, 2 Japanese, 2 Portuguese, 2 Mandarin, 1 Kosovan, 1 Somali, 1 Ethiopian, 1 Welsh, 1 Yoruba, 1 Jamaican Patois, 1 Polish, 1 Zulu, 1 Romanian, 1 Oshiwambo, 1 Vietnamese, 1 Hebrew, 1 Irish, 1 Gujarati, 1 Punjabi, 1 Urdu, 1 Singalese, 1 Hindi, and 1 Scots.

[13] One of the adult participants reported being deaf in one ear.

tested individually by a researcher in their school. The schools' head teachers agreed for their school to participate in the study before informed opt-out consent was obtained from parents/guardians of all participating children, and each child provided verbal consent before participating. Adults were tested online using the Gorilla experimental interface (www.gorilla.sc, Anwyl-Irvine et al., 2019), through which informed consent was also obtained. Participants were randomly assigned to one of the counterbalanced versions of the two experimental conditions (as in Experiment 1 and 2), though due to a technical error 22 of the final set of adult participants were given the LV condition and 26 were given the HV condition. In return for participating, adults received a payment of £3.50 (at a rate equivalent to £7.50 per hour), while children received stickers during the experiment and a certificate after completing the session.

Stimuli

Auditory stimuli items were identical to those used in Experiment 2 (see Table 14 above). Visual stimuli were identical to those used in Experiment 1 (see Figure 2).

Design

The experimental design was identical to that of Experiment 2. There were once more two variability conditions (HV and LV), and counterbalancing was identical to Experiment 1 and 2.

Procedure

The procedure for the training task as well as the pre/post-test tasks for Experiment 3 was identical to that of Experiment 2, except that the picture stimuli from

Experiment 2 were replaced by the shape stimuli from Experiment 1. All tasks for children were run using PsychoPy (Peirce, 2007) on laptop computers in quiet rooms at the school. For adults, all tasks were run using Gorilla (www.gorilla.sc, Anwyl-Irvine et al., 2019). Adult data was collected on 17 August 2018. Adult participants were able to run the experiment on a desktop or laptop computer, or a tablet device, and were told to do the experiment in a quiet environment. Stimuli were presented binaurally over headphones at a comfortable listening level.

Analyses

The approach for this experiment was as for the previous experiments. The same hypotheses as for Experiment 2 were tested, as well as additional hypotheses comparing performance in the two experiments. Note that adults, to make the fairest comparison, are compared to the online participants from Experiment 2. . As before, for completeness sake, all effects related to the hypotheses described above will be reported, but not all will be interpreted.

3.4.2 Results

Adult data

Discrimination task

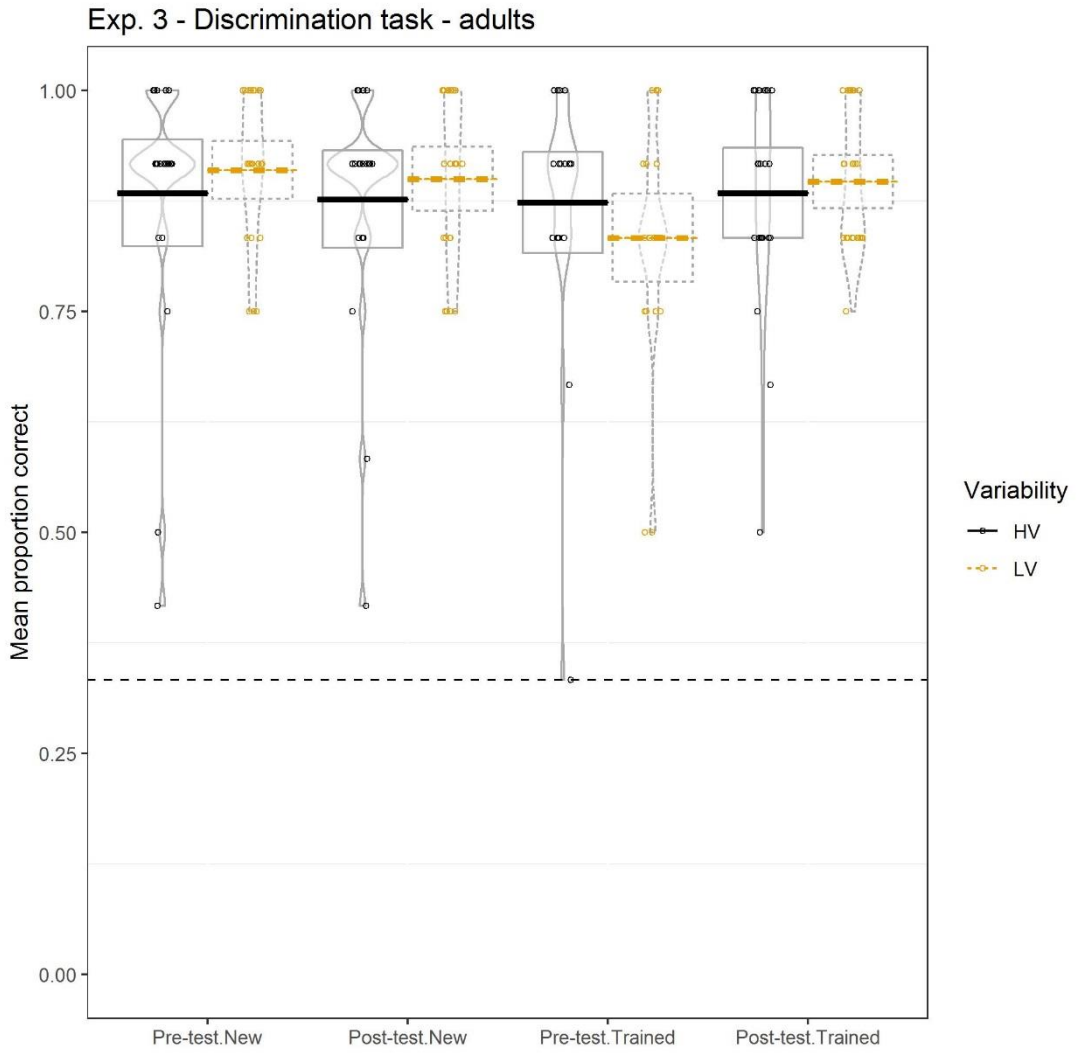


Figure 25. Pirate plot displaying accuracy results for adults on the pre- and post-test discrimination task of Experiment 3, comparing accuracy for new versus trained items as well as HV versus LV. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As in previous experiments, there was substantial evidence that adults are above chance, as shown in Figure 25 and Table 30. There was substantial evidence against adults improving from pre- to post-test, and against more improvement in trained

than untrained items. The evidence for more improvement in the HV over LV condition was ambiguous, as was the evidence for greater improvement in HV for untrained items. Note that it is not interesting to compare performance on the Discrimination task to that in Experiment 2, as there is no evidence of learning in either study.

Final structure Exp3_Adult_DM model: accuracy ~ session * condition * itemnovelty + speaker + (session*itemnovelty|participant) + (session|contrast)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	3.449	0.294	11.747	<.001	3.316	$9.50*10^{28}$	[0.1823, >3.8223]
improvement from pre- to post- test	session	0.004	0.219	0.018	.986	1.539	0.141	[0.6423, ∞]
greater improvement for trained items	item-novelty : session	0.523	0.354	1.478	.139	1.539	0.102	[0.4523, ∞]
greater overall improvement for HV training	condition: session	-0.257	0.291	-0.884	.377	1.539	1.169	[0, >38223]
greater improvement in HV greater for untrained items	*condition : item-novelty : session	0.617	0.576	1.072	.284	1.539	0.975	[0, >38223]

Table 30. Mixed model results for the Discrimination analysis of Experiment 3, for adults. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Training

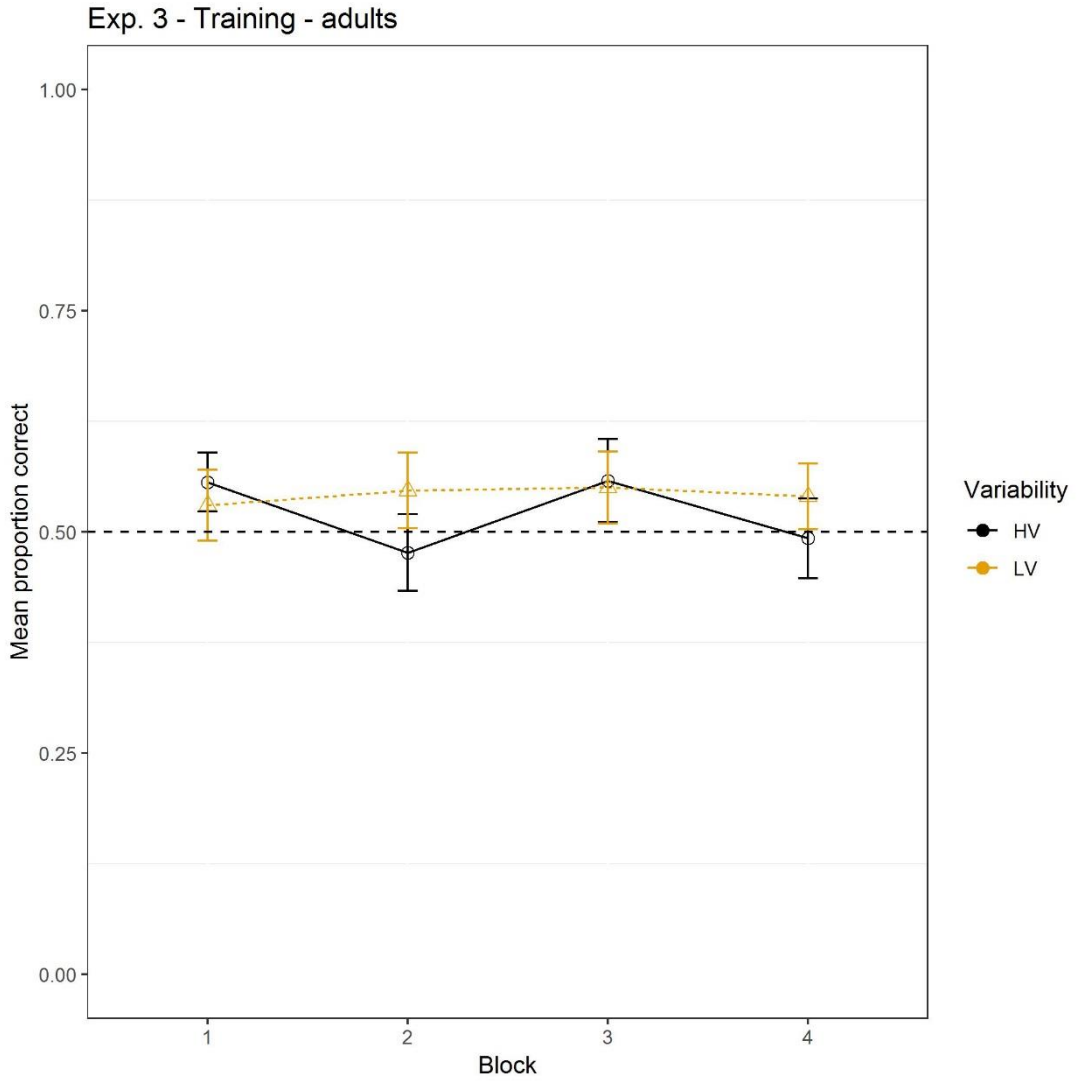


Figure 26. Adult accuracy scores in Training of Experiment 3 plotted over the 4 blocks, comparing HV and LV. Error bars show 95% CI, and the dashed line indicates chance level.

In contrast to Experiment 2, there was ambiguous evidence of learning in this experiment (as seen in Figure 26 and Table 31). The evidence for above chance performance was ambiguous, while there was substantial evidence against improvement across blocks. The evidence for there being greater improvement across blocks in LV than in HV was ambiguous. There was substantial evidence

that (online) participants were stronger in Experiment 2 than Experiment 3, both in general and in showing greater improvement across blocks (see Table 32).

Final structure Exp3_Adult_Train model: accuracy ~ block * condition + speaker + (block|participant) + (condition*block|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.134	0.083	1.616	.106	0.544	1.023	[0, 1.7117]
improvement across blocks	block	-0.013	0.030	-0.442	.659	0.07	0.288	[0.0617, ∞]
greater improvement in LV condition	*condition: block	0.060	0.065	-0.920	.358	0.131	0.99	[0, 0.4717]

Table 31. Mixed model results for the Training analysis of Experiment 3, for adults. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Experiment comparison

Final structure ExpComp_Adult_Train model: accuracy ~ block * condition * experiment + speaker + (block|participant) + (condition*block||word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Experiment2 participants outperform Experiment3 overall	experiment	0.625	0.069	9.095	<.001	0.195	$4.48*10^{15}$	[0.0417, >4.5517]
Experiment2 participants show greater improvement across blocks	experiment :block	0.210	0.042	5.010	<.001	0.122	48447.35	[0.0417, >4.5517]

Table 32 Mixed model results for the Experiment comparison for Training in adults.

Figure 27 below shows the degree of variability in the patterns of change over blocks across participants. As can be seen, there is quite a lot of individual variability: some adults show a deterioration, while others show improvement, or are stable across blocks.

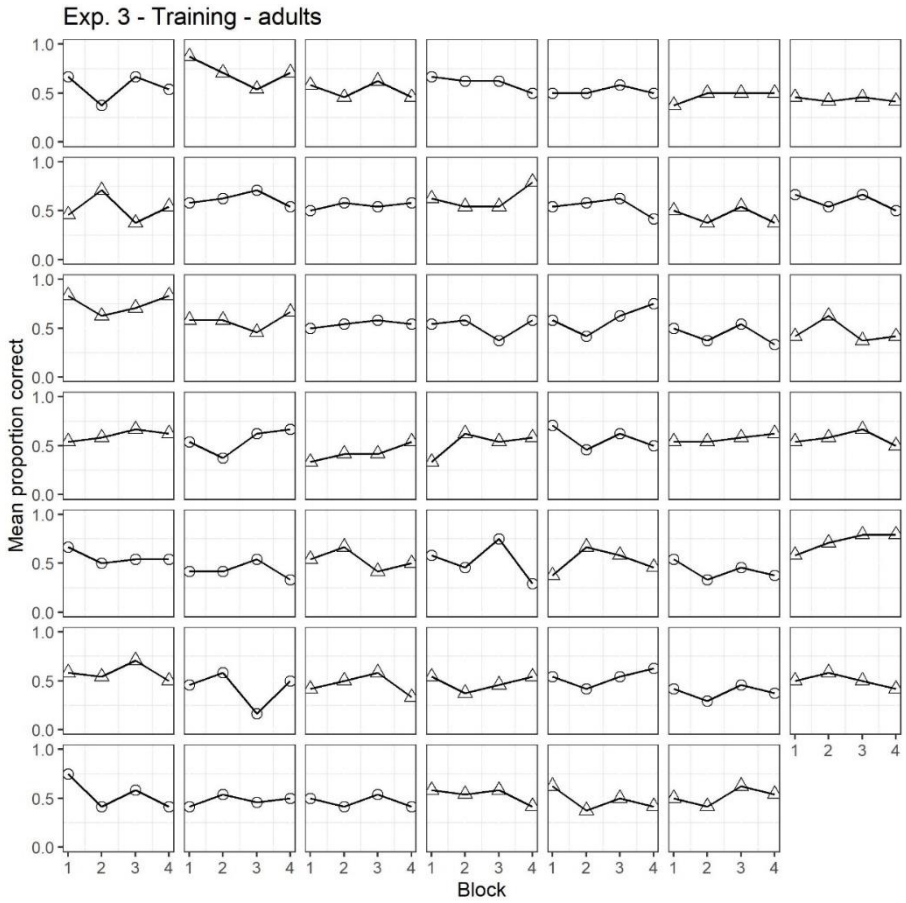


Figure 27. Adult accuracy scores in training of Experiment 3 plotted over the 4 blocks, split out by participant. Plots with circles indicate a participant received HV training, while triangles indicate LV training.

Figure 28 shows the pattern of change across the training blocks split out by vowel: while performance on <ui> is lower than that for <au> and <eu>, the overall pattern of change (and lack thereof) is similar.

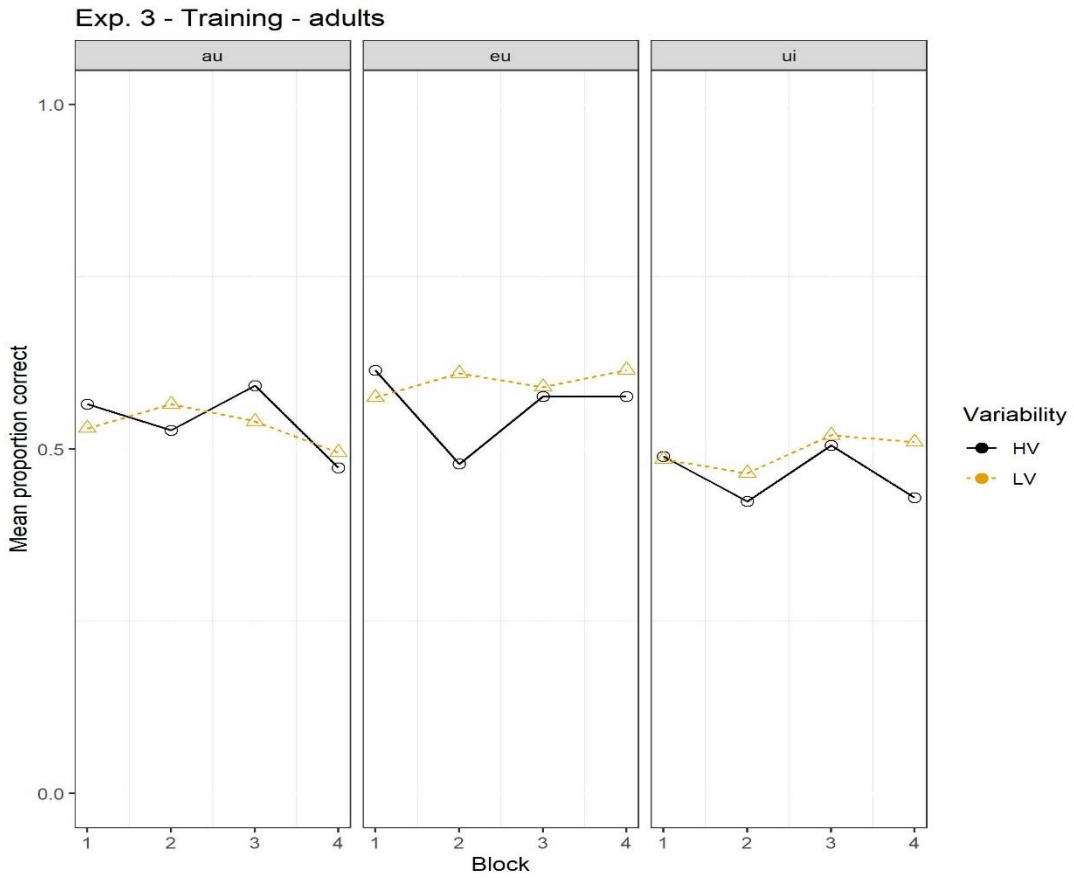


Figure 28. Adult accuracy scores in training of Experiment 3 plotted over the 4 blocks, split out by vowel, comparing HV and LV input.

Identification task (post-test only)

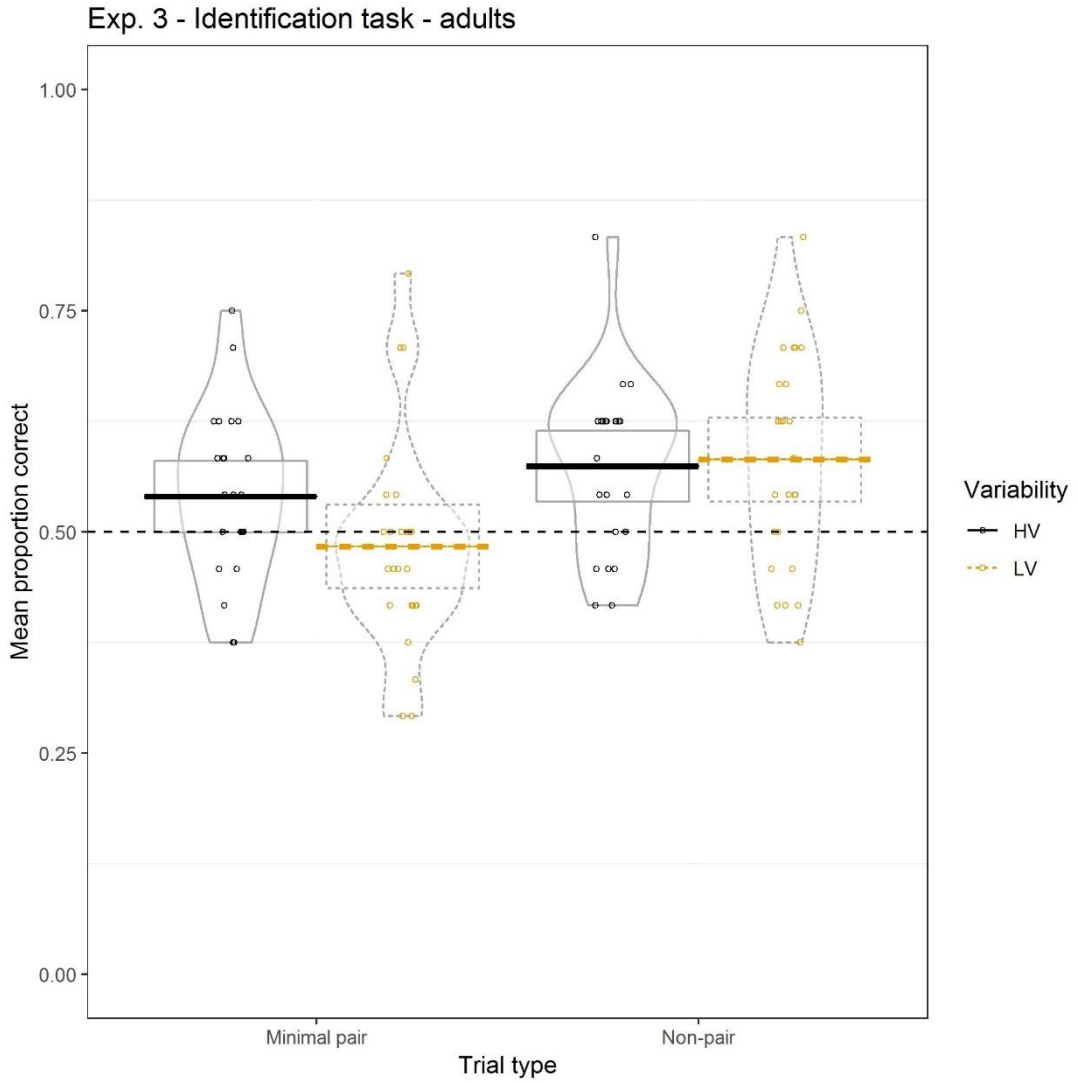


Figure 29. Pirate plots depicting adult accuracy results on the identification task of Experiment 3, comparing items presented in trained and novel items in the two variability conditions. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

For the Identification task, in contrast to Experiment 2, (as seen in Figure 29, Table 33, and Table 34) there was substantial evidence for above chance performance only in the non-pair items, with substantial evidence against above chance performance in the minimal pair items. Minimal pairs showed ambiguous evidence for greater performance in HV than LV, while non-pairs show substantial evidence

for the null. There was substantial evidence that (online) participants were stronger in Experiment 2 than Experiment 3 for both minimal pairs and non-pairs (see Table 35).

Minimal pairs:

Final structure Exp3_Adult_ID_minpair model: accuracy ~ condition + speaker + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.046	0.138	0.333	.739	0.584	0.306	[0.5334, ∞]
greater performance in HV condition	condition	0.239	0.131	1.688	.091	0.046	1.519	[0, 3.3134]

Table 33. Mixed model results for the minimal pair Identification analysis of Experiment 3, for adults.

Non minimal pairs

Final structure Exp3_Adult_ID_nonpair model: accuracy ~ condition + speaker + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.324	0.095	3.419	.001	3.661	17.8	[0.0834, >3.8434]
greater performance in HV condition	condition	-0.049	0.146	-0.338	.735	0.324	0.328	[0.3234, ∞]

Table 34. Mixed model results for the minimal pair Identification analysis of Experiment 3, for adults.

Experiment comparison

Final structure ExpComp_Adult_ID_minpair and ExpComp_Adult_ID_nonpair models:
accuracy ~ condition + experiment + speaker + (condition|participant) +
(condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Experiment 2 participants show greater performance than Experiment 3 - minimal pairs	Experiment	0.716	0.133	5.372	<.001	0.189	17560.69	[0.0834, >3.8434]
Experiment 2 participants show greater performance than Experiment 3 - non minimal pairs	experiment	3.658	0.245	14.910	<.001	1.368	$2.08*10^{46}$	[0.0834, >3.8434]

Table 35. Mixed model results for the Experiment comparison for the Identification analysis of Experiment 2 and 3, for adults.

Child data

Discrimination task

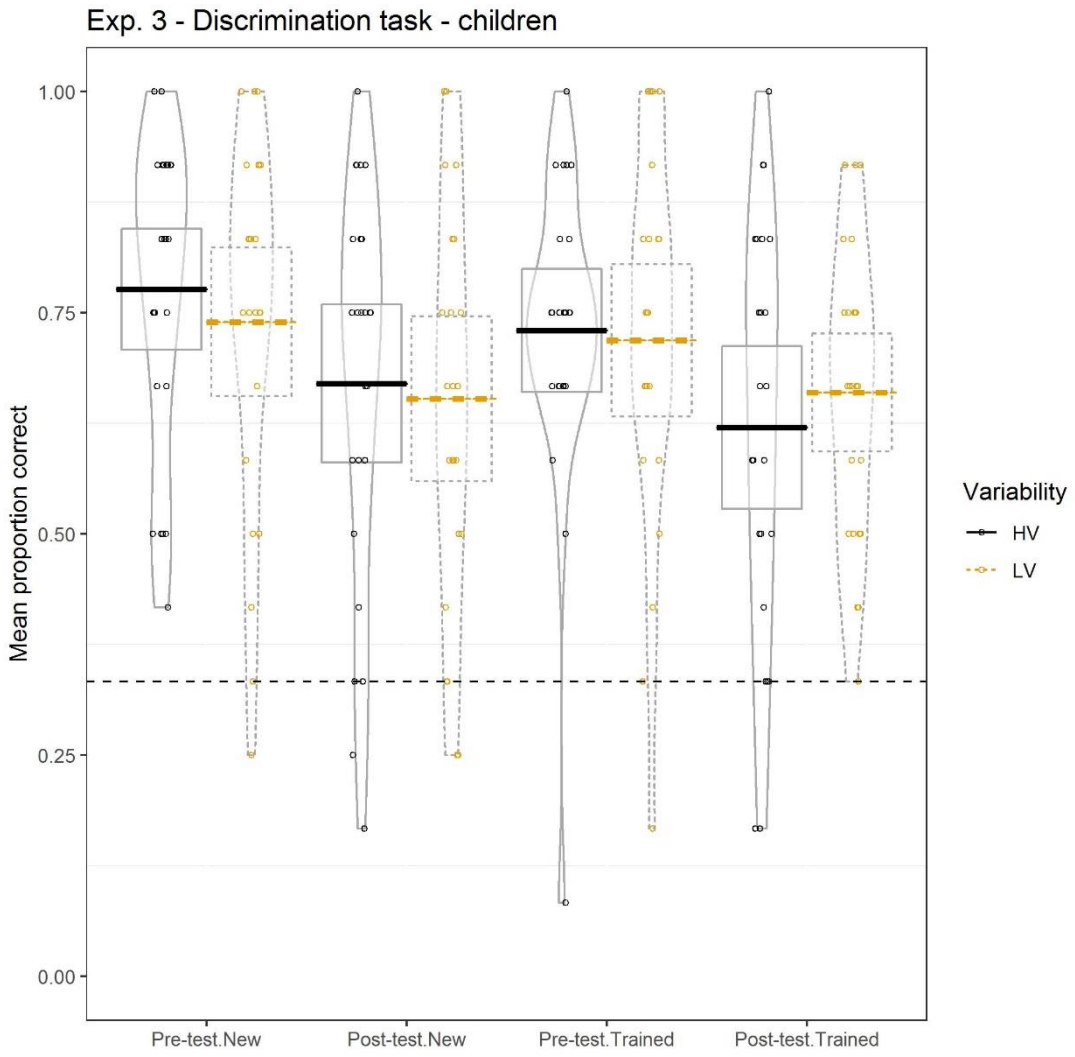


Figure 30. Accuracy results for children on the pre- and post-test discrimination task of Experiment 3, comparing accuracy for new versus trained items as well as HV versus LV. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

There was substantial evidence that children are above chance in Discrimination, as shown in Figure 30 and Table 36. However, there was substantial evidence against improvement from pre- to post-test, against greater improvement in the HV than LV condition and vice versa, and against greater improvement on non-

pair items than on minimal pair items. There was also substantial evidence against greater improvement in HV for untrained items. Note that as for adults, it is not interesting to compare performance on the Discrimination task to that in Experiment 2, as there is no evidence of learning.

Final structure Exp3_Child_DM model: accuracy ~ session * condition * itemnovelty + speaker + (session*itemnovelty|participant) + (session|contrast)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	1.715	0.170	10.115	<.001	1.983	$1.94*10^{21}$	[0.1823, >3.8223]
improvement from pre- to post- test	session	-0.536	0.139	-3.863	<.001	1.665	0.014	[0.1823, ∞]
greater improvement for trained items	item-novelty : session	0.106	0.213	0.500	.617	1.665	0.196	[0.9723, ∞]
greater overall improvement for HV training	condition: session	-0.187	0.268	-0.699	.485	1.665	0.097	[0.4523, ∞]
greater overall improvement for LV training	*condition: session	0.187	0.268	0.699	.485	1.665	0.303	[1.5123, ∞]
greater improvement in HV greater for untrained items	*condition : item-novelty : session	0.115	0.401	0.286	.775	1.665	0.295	[1.4823, ∞]

Table 36. Mixed model results for the Discrimination analysis of Experiment 3, for children. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Training

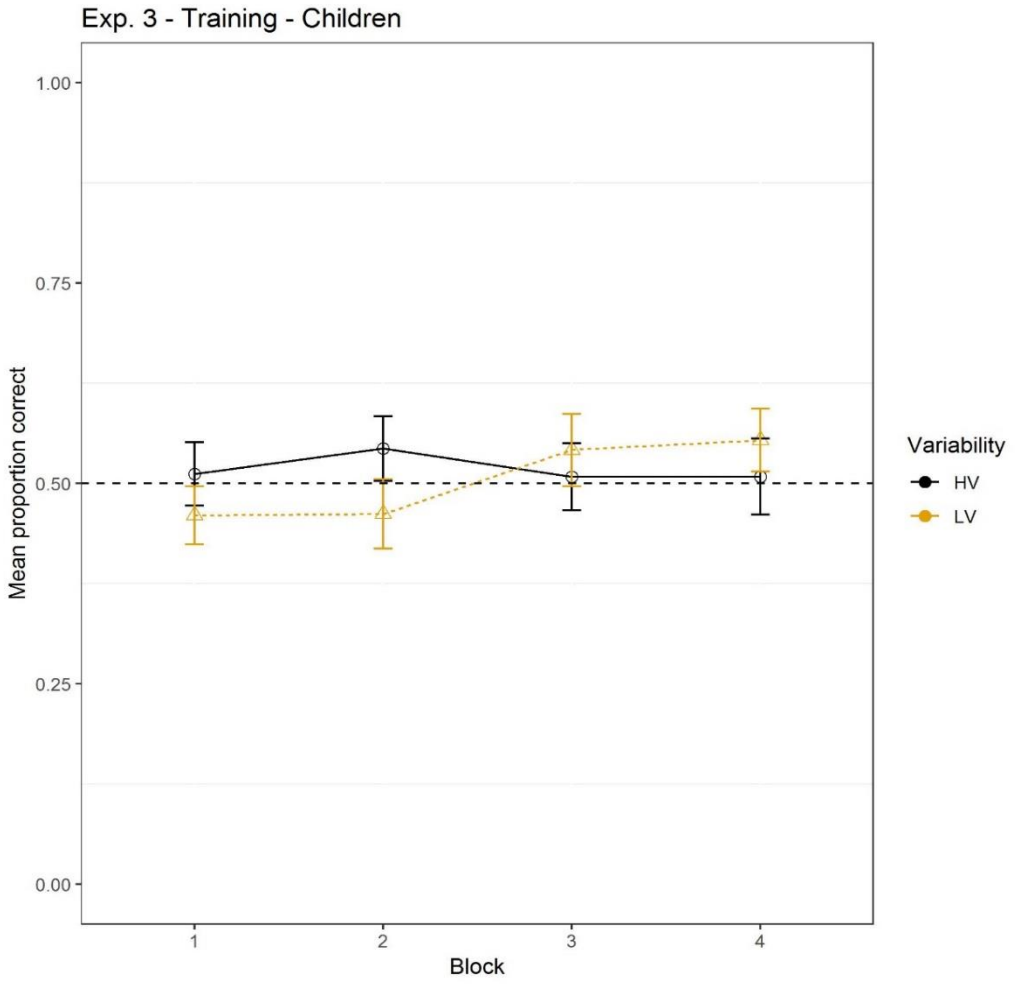


Figure 31. Child accuracy scores in training of Experiment 3 plotted over the 4 blocks, comparing HV and LV. Error bars show 95% CI, and the dashed line indicates chance level.

As can be seen in Figure 31 and Table 37, the evidence was ambiguous as to children being above chance overall, as well as for them improving across blocks. There was substantial evidence against there being greater improvement in LV, and when breaking this down there was evidence against improvement over blocks for LV but ambiguous evidence for HV. There was substantial evidence for participants being stronger in Experiment 2 than Experiment 3 and ambiguous evidence for them showing greater improvement across blocks in Experiment 2 than Experiment 3 (see Table 38).

Final structure Exp3_Child_Train model: accuracy ~ block * condition + speaker + (block|participant) + (condition*block|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.044	0.032	1.370	.171	0.359	0.407	[0.4517, >4.5517]
improvement across blocks	block	0.061	0.033	1.814	.070	0.083	2.929	[0.0917, 1.0517]
greater improvement in LV condition	*condition: block	0.164	0.054	3.066	.002	0.037	0.309	[0.0417, ∞]
FOLLOW-UP Improvement across blocks in LV	Block: conditionLV	0.150	0.043	3.501	<.001	0.045	33.864	[0.04, >4.595]
FOLLOW-UP Improvement across blocks in HV	Block: conditionHV	-0.011	0.042	-0.252	.801	0.045	0.796	[0, 0.149]

Table 37. Mixed model results for the Training analysis of Experiment 3, for children. *The sign of this fixed effect is changed from the model coding to be in line with the hypothesis.

Experiment comparison

Final structure ExpComp_Child_Train model: accuracy ~ block * condition * experiment + speaker + (block|participant) + (condition*block||word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Experiment 2 participants outperform Experiment 3 overall	experiment	0.309	0.043	7.197	<.001	0.195	$2.29*10^{10}$	[0.0417, >4.5517]
Experiment 2 participants show greater improvement across blocks	Block: experiment	0.018	0.038	0.489	.625	0.122	0.445	[0, 0.1617]

Table 38. Mixed model results for the Experiment comparison for Training in children.

Figure 32 below shows the degree of variability in the patterns of change over blocks across participants. As can be seen, the majority of children show a similar

lack of change across blocks, although some children do seem to show a degree of improvement.

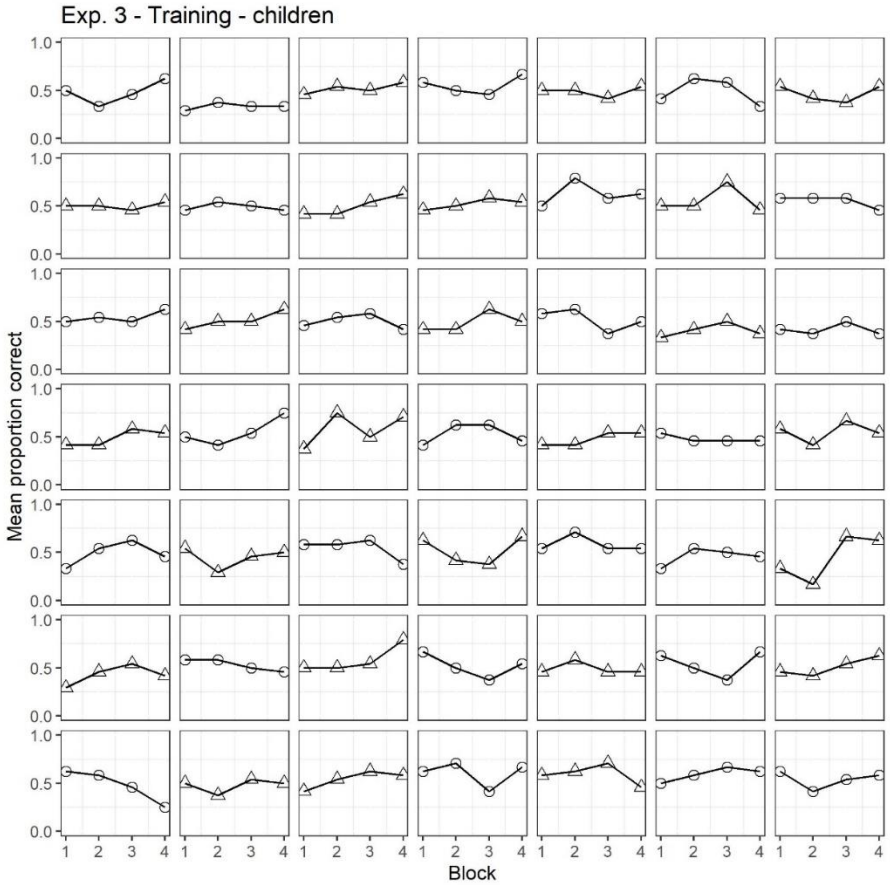


Figure 32. Child accuracy scores in training of Experiment 3 plotted over the 4 blocks, split out by participant. Plots with circles indicate a participant received HV training, while triangles indicate LV training.

Figure 33 shows the pattern of change across the training blocks split out by vowel: performance on the three vowels is similar, although <au> shows less improvement than <eu> and <ui>.

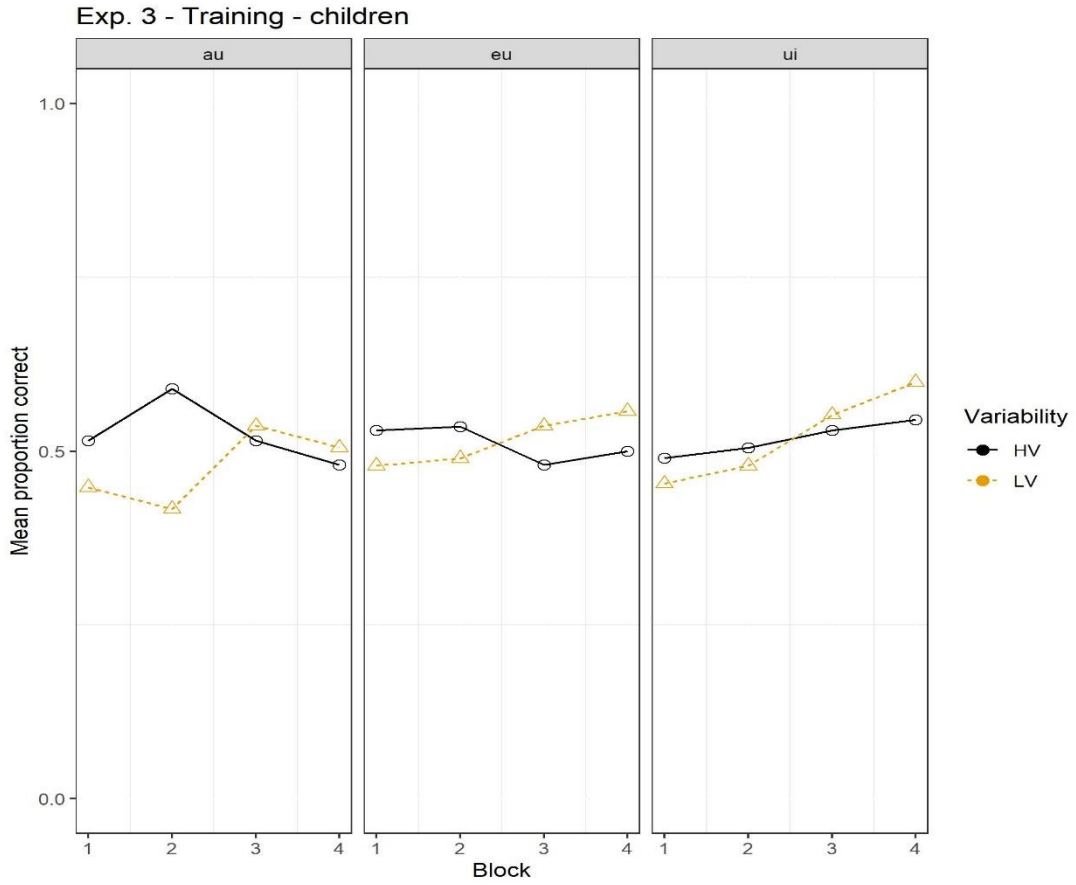


Figure 33. Child accuracy scores in training of Experiment 3 plotted over the 4 blocks, split out by vowel, comparing HV and LV input.

Identification task (post-test only)

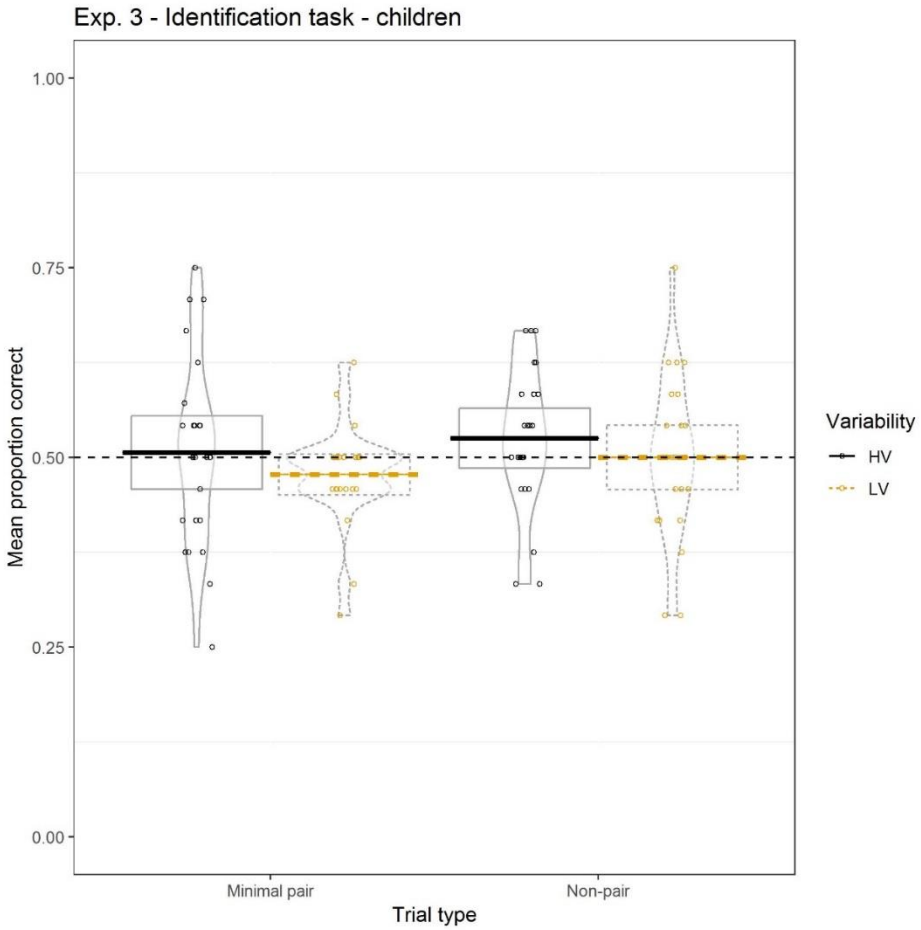


Figure 34. Child accuracy results on the identification task of Experiment 3, comparing items presented in minimal pairs and non-pairs in the two variability conditions. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 34, Table 39, and Table 40, there was substantial evidence against children performing above chance on both the minimal pair and the non-pair trials of the Identification task. There was ambiguous evidence for HV outperforming LV in minimal pairs, while there was substantial evidence for the null in the non-pairs. There was substantial evidence against LV outperforming HV for both minimal pairs and non-pair trials. There was substantial evidence for participants being stronger in the non-pairs for Experiment 2 than Experiment 3, but substantial evidence for the null for the minimal pairs (see Table 41).

Minimal pairs

Final structure Exp3_Child_ID_minpair model: accuracy ~ condition + speaker + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	-0.036	0.066	-0.543	.587	0.298	0.146	[0.1334, ∞]
greater performance in HV condition	condition	0.108	0.131	0.824	.410	0.298	0.827	[0, 0.8534]
greater performance in LV condition	condition	-0.108	0.131	-0.824	.410	0.298	0.239	[0.2134, ∞]

Table 39. Mixed model results for the minimal pair Identification analysis of Experiment 3, for children.

Non minimal pairs

Final structure Exp3_Child_ID_nonpair model: accuracy ~ condition + speaker + (condition|participant) + (1|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.050	0.060	0.826	.409	1.899	0.073	[0.4034, ∞]
greater performance in HV condition	condition	0.103	0.120	0.853	.394	1.899	0.148	[0.8234, ∞]
greater performance in LV condition	condition	-0.103	0.120	-0.853	.394	1.899	0.038	[0.1834, ∞]

Table 40. Mixed model results for the non-minimal pair Identification analysis of Experiment 3, for children.

Experiment comparison

Final structure ExpComp_Child_ID_minpair model: accuracy ~ condition + experiment + speaker + (condition|participant) + (1|word)

Final structure ExpComp_Child_ID_nonpair model: accuracy ~ condition + experiment + speaker + (condition|participant) + (condition|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Experiment 2 participants show greater performance than Experiment 3 - minimal pairs	intercept (min pairs) - chance	0.328	0.086	3.813	<.001	0.137	0.12	[0.0834, ∞]
Experiment 2 participants show greater performance than Experiment 3 - non minimal pairs	intercept (non-mp pairs) - chance	1.541	0.104	14.770	<.001	0.301	$1.18*10^{42}$	[0.0834, >3.8434]

Table 41. Mixed model results for the Experiment comparison of Experiment 2 and 3 for Identification in children.

Comparing adults and children

Once more, the analyses reported above suggest some differences between adults and children. As before, wherever an effect was found in at least one age group, it is of interest to compare the adults and children. Results are in Table 42.

There was substantial evidence for adults outperforming children in the Discrimination task. There was substantial evidence for adults outperforming children on the non-pairs of the Identification task, while evidence was ambiguous for the minimal pairs.

Final structure Exp3_AgeComp_DM model: accuracy ~ session * condition * itemnovelty + group + speaker + (session*itemnovelty|participant) + (session|contrast)

Final structure Exp3_AgeComp_Train model: accuracy ~ block * condition * group + speaker + (block|participant) + (group.ct + condition.ct:block.ct|word)

Final structure Exp3_AgeComp_ID_minpair model: accuracy ~ condition * speaker + group + (condition|participant) + (condition|word)

Final structure Exp3_AgeComp_ID_nonpair model: accuracy ~ condition * speaker + group + (1|participant) + (1|word)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
adults outperform 7 years olds in discrimination	Discrimination model: *age*	1.448	0.200	7.230	<.001	1.204	$3.63*10^{10}$	[0.1823, >3.8223]
Adults outperform children in training task.	Training model: *age*	0.086	0.080	1.077	.281	0.186	1.077	[0, 0.7217]
Adults show greater performance than children in identification task – minimal pairs.	Identification model: *age*	0.077	0.088	0.878	.380	0.251	0.743	[0, 0.6134]
Adults show greater performance than children in identification task – non-minimal pairs.	Identification model: *age*	0.267	0.087	3.076	.002	1.441	13.382	[0.0834, >3.8434]

Table 42. Mixed model results for the Adults versus Children Comparison of Experiment 3.

3.4.3 Discussion

In the third experiment, to establish whether the learning found in Experiment 2 was caused by the use of meaningful pictures or the switch from 3AFC to 2AFC, Experiment 2 was repeated but with a 2AFC symbol mapping task. Input variability

was still the key manipulation, and performance was again measured through means of a discrimination and identification task.

Since adults showed evidence of learning during Training and in Identification in both Experiment 1 and Experiment 2, the key focus here is on the children's performance. In Experiment 3, in Training, the evidence that performance was above chance was ambiguous, as was the evidence for improvement across blocks. However, critically, there was substantial evidence that overall performance was weaker than in Experiment 2, though the evidence that improvement across blocks was stronger in Experiment 2 was ambiguous. In Identification, there was substantial evidence against above chance performance for children for both minimal pairs and non-pairs (and thus not surprisingly, either substantial evidence for the null or ambiguous evidence for the hypotheses concerning variability). Children are thus not able to generalise in this paradigm, likely because there was very little if any learning at all in training. Consistent with this, again, there was substantial evidence that performance in Experiment 2 – where they *did* generalise – was stronger for the non-pairs, though this was not substantial for the minimal pairs.

Interestingly, there was also evidence for adults that, like children, they had more difficulty in this experiment than in Experiment 2. In Training, there was substantial evidence that both overall learning and steepness of the learning slope was weaker in this experiment, and in fact the evidence for above chance performance in Training here was ambiguous, with substantial evidence against

improvement across blocks. In the Identification task, there was substantial evidence against above chance performance for minimal pairs, but substantial evidence for above chance performance on non-pairs, suggesting some learning of the associations between the shapes and the words, but poor learning of the phonetic contrast when asked to generalise knowledge from Training to novel speakers. Importantly, for both trial types, there was substantial evidence that adults in Experiment 2 – where evidence for learning and generalisation *was* found – outperformed adults in Experiment 3. In terms of age differences, performance was not compared where neither group showed evidence of above chance performance or learning. However, note that adults again outperformed children on Discrimination and the non-pairs of the Identification task.

As in the previous two experiments, in Discrimination there was evidence for above chance performance in both age groups, with adults outperforming children, but no learning as a result of Training in either group. This consistent finding will be revisited in the general discussion in Section 3.5.

Overall, this experiment clearly showed that the differences between Experiment 1 and Experiment 2 were not just due to lower task demands of having a 2AFC over a 3AFC training task. Instead, it can be seen here that both adults and children are capable of doing a task in which they create mappings between words and pictures that depend on being able to hear a non-native vowel contrast. However, both children and adults have more difficulty in learning mappings where they are asked to associate abstract symbols with those non-native vowels within the

stimuli, despite being given identical auditory stimuli and using the same 2AFC task.

3.5 General discussion

This study aimed to investigate whether the HVPT paradigm is useful for L2 speech sound learning in children and, if so, whether they show the same HV advantage as adults have been shown to do. However, after seeing very weak learning in Experiment 1, especially for children, it was hard to draw conclusions about variability effects. Therefore, the aim for Experiment 2 was to develop a task that would promote learning. Increased learning was indeed achieved, but still no evidence of a variability effect was found. Experiment 3 followed up on the differences in learning found in the two earlier experiments, showing that greater learning and generalisation were due to the use of a more concrete picture-based training task rather than due to procedural differences between the use of a 2AFC and 3AFC training task. The remainder of this discussion will first focus on the original question regarding the role of variability in non-native speech learning for children. The differences in the use of an orthography-based paradigm versus a picture-based paradigm will then be discussed, before considering age differences in the results. The discussion will finally consider a surprising but consistent finding across the three experiments: that Training did not change performance in a Discrimination task.

3.5.1 Variability

Starting with the original question of variability, there was no effect of training variability in adults or children for any of the three experiments discussed above. This is contrary to what has been found in the literature discussed in Section 1.3.5, and goes against the theoretical idea that having encountered multiple talkers in your training input helps you generalise to novel talkers, due to having encountered variability across voices before, making it easier to ignore idiosyncratic features when understanding a new voice. Other studies have also found this HV benefit, although there are actually relatively few directly comparing HV and LV input (as discussed in Chapter 1.3.5). Due to publication bias, it is quite possible that null findings have not been published (Ferguson & Heene, 2012).

Nevertheless, it is important to consider whether there is anything in particular about the current study that could explain why no HV benefit can be seen. First, in Discrimination it is hard to interpret any findings with regard to variability, as no effect of learning was found overall. There are also some places where no learning was seen in the Identification task (i.e. children in Experiment 1 and 3, and adults with minimal pair items in Experiment 3). However, even in the places where learning was found (i.e. lab-based adults, online adults, and children in Experiment 2, as well as non-pairs in adults for Experiment 3), there was still no evidence for an HV benefit. This is unexpected, since the post-test Identification

task is where previous studies have found a benefit of HV training input (Sadakata & McQueen, 2013). What is the reason that this effect is not seen here?

As briefly touched upon earlier, one possibility is that this is due to the single-session nature of the experiments: most HVPT studies have been multiple-session training studies, often lasting at least 5 sessions (see Lengeris & Hazan (2010) and Iverson & Evans (2009)) or ten to fifteen sessions (see Logan et al. (1991), Lively et al. (1993), Shinohara (2014), Giannakopoulou et al. (2010)) if not more: Yamada (1993) even trained up to 45 sessions. Most of the learning seems to occur in the first 10 sessions (Logan & Pruitt, 1995), so the single training session used in this experiment might not have been enough exposure to show a difference in variability conditions. That said, accent training studies (Clopper & Pisoni, 2004) and vocabulary training studies (Barcroft & Sommers (2005); Sinkevičiūtė et al. (2019)) which have compared similar HV/ LV manipulations generally have been single-session training studies as well, and they do find variability effects. However, this could be due to a slightly different focus for the learners: in the accent and vocabulary studies, learners are either memorising whole words, or relying on their knowledge of particular identifiable accent features in a language they are familiar with. For the phonetic training studies, learners need to adjust their existing phoneme categories to be able to differentiate the non-native categories that are used. It might be the case that in a short study such as this one, rather than changing their existing L1 phoneme categories, learners are relying on temporary phoneme categories created in the context of the experiment, as

suggested by Heeren & Schouten (2008). Having multiple training sessions would allow for a potential longer term shift in the learners' phoneme categories, rather than them relying on temporary categories used for the duration of the experiment only. It might be that when this shift occurs, a HV benefit would emerge, as more variation in the input should help in more quickly adjusting the phoneme categories to the present vowels.

Another factor that is influenced by the one-session nature of the paradigm used in the current study, is that it left no possibility for sleep consolidation. Fuhrmeister & Myers (2017) found that introducing variability at test, in the form of novel items or speakers, reduced performance on trained items as it seemed to destabilise learning, and similar phonetic training studies from that lab (e.g. Earle et al. (2017); Earle and Myers (2015a,b)) have found sleep consolidation to be beneficial to both training and generalisation performance. It could be the case that sleep consolidation is needed for any variability benefit to unfold in learning. However, again, in studies involving vocabulary learning, one session seems to be enough to learn non-native words without the need of sleep consolidation, so this may not be a complete account.

In addition to not finding the hypothesised HV benefit, this study also found no evidence for an LV benefit in training, nor for an LV benefit at test for child participants. Recall that Evans & Martín-Alvarez (2016) and Giannakopoulou et al. (2017) did find a benefit of LV in training using a similar task with10 training sessions. Finding such LV benefit in training would indicate that the training

materials were easier in the LV condition with one talker throughout the training session, compared to the HV condition where adapting to multiple talkers requires more cognitive effort. The data from this study shows no clear LV benefit, although note that evidence for this effect was ambiguous in most of the data sets apart from in the Online adults from Experiment 2 where there is evidence for H1 and the children in Experiment 3 where there is evidence for the null, making it impossible to draw strong conclusions. Tentatively, it seems that participants are not finding the HV much harder than the LV materials. This could be due to a procedural difference: Giannakopoulou et al. (2017) alternated speakers on a trial-by-trial basis, resulting in a higher task demand, while the current study blocked the speakers in line with most previous HVPT paradigms, as suggested by Logan & Pruitt (1995). This additional task demand in Giannakopoulou et al. (2017) could have been the push for the LV benefit to emerge in training, while this was not seen in the current study (though see Dong et al. (2019), where LV outperformed blocked HV input, and both outperformed unblocked HV input over multiple training sessions). Alternatively, or in addition, it could be that seeing the LV advantage requires time to "adapt" to the talker. Finally, since no LV advantage was found in training, there is also no reason to see it for children at test as Giannakopoulou et al. (2017) did.

3.5.2 Orthography vs pictures

Although not part of the original research questions, perhaps the most interesting finding of this study is the lower performance on the paradigm using an

orthography-like symbol system compared to the word learning-like picture mapping paradigm. Why might it be so much harder when participants have to pick out the individual vowels in order to learn the mappings?

Remember that participants are theoretically learning more mappings in the picture-based task, where they are learning 12 individual pictures, compared to the symbol-based task, where they are just learning 3. This suggests that it is something about having to map the individual phonemes to the symbols that makes the task inherently harder, rather than the number of mappings they are making. It seems that picking out individual phonemes from the rest of the segments is hard, presumably because phonemes can be very different in different contexts due to connected speech processes such as coarticulation (see Chapter 2). This is a known hurdle when children are learning to read in their native language, where phonological working memory (Gathercole & Baddeley, 1993) and phonological awareness (Snowling & Hulme, 1994) seem to play an important role in the early stages of learning to read, correlating with reading performance later on. Phonological working memory refers to a component of memory specifically used to maintain spoken material while processing speech, while phonological awareness is the metalinguistic awareness of the phonological representations underlying speech, that is, the ability to reflect on the phoneme structure of spoken words. When these abilities are still developing, it might be harder to store spoken information long enough to fully process it, and to break down the input into the underlying phoneme structure in order to map individual phonemes to

orthographic symbols (see Gathercole (1999)). If this phoneme-to-symbol mapping is already so difficult and reliant on phonological awareness in children's L1, it must be even harder for non-native contrasts where there is the added complication of needing to learn to hear the difference between the phonemes before being able to segment them and map them to symbols consistently. Note that even adults, who have a more developed phonological memory, found the orthography-like Experiments 1 and 3 harder than Experiment 2, suggesting phonological awareness is not the only factor in the low performance.

How does this difficulty mapping novel orthography to vowel sounds fit with findings from previous studies with children which *have* successfully used orthography or orthography-like symbols in HVPT training? Wang & Kuhl (2003) use a paradigm perhaps most similar to this study, where they train adults and children to map lexical tones to specific animals (one per tone, akin to the one symbol per vowel here). However, it is possible that as they are suprasegmentals, it is easier to separate the tones out from the rest of the word than it was for the vowels in the current set of experiments. Another obvious difference was again the length of training – that study used 6 training sessions while the current study used one session. Other studies in the literature which have used orthography in HVPT with children have also had multiple training sessions between pre- and post-test (Giannakopoulou et al., 2010, 2013; Heeren & Schouten, 2010; Shinohara, 2014; Shinohara & Iverson, 2013). Another factor to consider is that in many of these previous studies, learners tend to have at least some familiarity with the

orthography that is used already. This could be because the contrasts that are trained on are consonants that map consistently onto a specific spelling (e.g. Lively et al. (1993); Logan et al. (1991); Sadakata & McQueen (2013); Shinohara & Iverson (2013)), rather than vowels which tend to have more variable spelling. When vowels are used in training, learners tend to already have some basic experience of the L2 they are learning. In the current study, where learners have no previous experience of the L2, the real Dutch spelling could not be used since children would be expected to link these to the phonology of their L1 (English), leading to unwanted negative transfer from the L1. This might be especially relevant in child learners as in addition to having had less L2 exposure, their L1 spelling-to-sound mappings might not be stable yet and they might still associate a multitude of spellings, including incorrect ones, with specific vowels. For them, seeing the orthography for novel non-native contrasts might have led to more unexpected transfer: where in adults the small set of potentially confused native vowels can be identified, for children many more options might be realistic candidates if their L1 orthographic awareness has not stabilised yet.

In addition to familiarity to the orthography, it could also be the case that there is some additional information in the orthography that helps learners improve on the non-native contrasts, i.e. the orthography might provide visual cues to the identity of the contrast. This can be seen with adults in for instance Sadakata & McQueen (2013), where the consonant length distinction between /s/ and /ss/ was represented through an additional letter, or with children in Giannakopoulou et

al. (2017) where words spelled with <ee> generally contained the longer vowel /i/, compared to <i> which mapped onto the short vowel /ɪ/.

A possible implication of the current study is that adults and children can learn vowel contrasts better in a word learning context than in an orthography context. However, it might be the case that orthography could be beneficial still when the learners are already familiar with some of the L2 orthography, or where the orthography of the L2 partly overlaps with the L1 leading to positive transfer (such as for consonants). The role of orthography is a question that will be further investigated in Chapter 4.

3.5.3 Age differences

Another question that has been of interest in the literature is that of age differences in language learning. Recall from Section 1.3.4 that previous studies have generally found that adults do better overall in these experiments, but this may be due to greater previous experience and thus a higher starting point in learning. In terms of clear evidence of *learning* due to training exposure, many studies suggest similar learning rates in training, but some have found that younger learners can outperform adults in measures of generalisation (Giannakopoulou et al., 2010, 2013; Shinohara, 2014). In the current study, despite age-groups having matched previous knowledge of the target L2 Dutch (i.e. none at all), all three experiments discussed above found that adults generally not only outperform children in all tasks, including at pre-test in Discrimination, in Experiment 2 they also show a

steeper learning curve during Training (with substantial evidence for this with the lab-based participants), and they show substantially stronger performance in the Identification task in all experiments. This result goes against the idea that younger learners should be better at acquiring an L2 than older learners, and contrasts with the results of Shinohara (2014); Shinohara & Iverson (2013) who found that one of their child learner groups outperformed adults in the amount of improvement from pre- to post-test, which they attributed to children having enhanced plasticity, as well as adults having increased L1 exposure which might interfere with their L2. Similarly, Giannakopoulou et al. (2013) found that children showed a much greater amount of post-training improvement compared to adults after being trained on a comparable HVPT task. In contrast, the results from the current experiments indicate adults might actually learn faster than children in the current paradigm.

However, note that while Shinohara (2014) found 8-12 year-olds outperformed adults, there was no such age benefit for their youngest age group (7-8 year-olds), who are similar in age to the children tested in the current study. They put down the lack of finding an age benefit in the youngest group to the use of unfamiliar orthography. Linking this to the current study, the use of unfamiliar orthography could explain the lack of such an age effect in Experiment 1 and 3, though note adults in the current study were naïve learners of the L2 as well and would have been a lot less familiar with the L2 orthography than the adults in Shinohara (2014) were. However, since no evidence of children outperforming adults was found in

Experiment 2 either, where orthography did not play a role, the current study seems to indicate that this cannot be an entire explanation of children's slower learning.

Additionally, there might be other factors at play in the paradigm that might aid adult learning while discouraging children. One crucial point to consider once more is the length of training: if the early stages of learning are mostly about getting used to the task, children might have more difficulty getting to grips with the task demands before actually getting to learning the contrasts. Thus adults can go faster in these early stages of learning, as seen in the steepness of the slope of improvement.

3.5.4 Discrimination

A remarkable, yet consistent finding across the experiments reported here is that while in some cases evidence was found of learning in Training, and of generalization in Identification, this learning never influenced the Discrimination results in post-test. It could be suggested that this might be due to the choice of discrimination task: all three experiments used oddity tasks, which are known to be cognitively more demanding than for example AX discrimination tasks (Logan & Pruitt, 1995b). However, this is in opposition with what has been found in Giannakopoulou et al. (2017, 2010) using a similar oddity task, where training improved the discrimination ability of both adults and children. The lack of a difference in variability conditions on the discrimination task is reminiscent of

Sadakata & McQueen (2013), who found that for a discrimination task improvement was similar for both HV and LV training input, but even there improvement from pre- to post-test was found for both conditions, contrary to the findings from the current study. A potential explanation for the lack of change in discrimination is offered by Heeren & Schouten (2010), who found a similar discrepancy between identification and discrimination results in their perceptual training task. They propose, based on Heeren & Schouten (2008), that in phonetic training experiments learners may rely on temporary phoneme categories, created in the context of the experiment, rather than change their existing phoneme categories. Learners might not be able to use these temporary categories in a discrimination task where they have to directly compare stimuli, while they could be adequate to successfully complete identification tasks. This seems like a plausible explanation for the results found here: it is very likely that the short amount of exposure in this study was not enough for learners to fully develop novel phoneme categories or adapt their existing phoneme categories to the novel contrast. They might instead have created temporary phoneme categories with which they were able to perform the Training and Identification task, but much more exposure would be required to show a change in their discrimination abilities.

3.5.5 Future research

On the whole, the current study raised a number of questions concerning the effect of input variability in phonetic training. One factor that seems important to

investigate, is lengthening the training from one session to multiple sessions to see if increasing the amount of exposure participants receive allows a difference between HV and LV training input to emerge. In addition, multiple sessions might allow enough time for a potential performance difference between adults and children to arise, as well as providing opportunity to investigate a longer term shift in phoneme categories manifesting itself in a change in learners' discrimination abilities. To investigate whether learners use temporary phoneme categories which they develop during training instead of changing their existing categories, as suggested by Heeren & Schouten (2010), another option is to give them the same post-test with a delay in between training and post-test. This delay should mean the temporary categories they built up are no longer maintained by the time they have to sit the delayed post-test, which should manifest in a different outcome in their identification task, but no change in their discrimination abilities since they were not using those temporary categories there to begin with. This effect of lengthening the training and adding a delay between training and post-test will be further investigated in Chapter 4.

Furthermore, to investigate whether learners are using more shallow methods in learning during these tasks, the pre- and post-tests could be adapted to try and pull apart what parts of the information provided in the input are being used by the learners. In order to succeed at the tasks and learn the picture-mappings, participants need to be able to tell apart the minimal pairs. This process appears to require being able to discriminate the non-native vowels, but it might be the case

that participants can rely on the context to perform the identification task, while not being able to discriminate the vowels at a lower individual segment level. Additionally, while the use of meaningful pictures turned out to be key in generating successful learning in the current paradigm, past research suggests that adding orthography – at least when L2 orthography is familiar – could potentially be beneficial for learners as well. The extent to which child participants can make use of pictures versus orthography will be investigated in Study 2, which attempts to tease apart how these different sources of information are used in phonetic training.

To conclude, the current study trained native English adults and 7-8 year-old children on a Dutch three-way vowel contrast. The initial paradigm using abstract symbols that mapped to phonemes proved to be hard for children, while an adapted picture-meaning mapping task promoted greater learning in both adults and children. A follow-up experiment confirmed that the increased learning was due to the use of meaningful pictures to create a word learning-like task, rather than a change in task demands. Finally, the current study found no evidence for an effect of input variability in training.

Chapter 4. Study 2

4.1 Introduction

The previous chapter found that contrary to predictions, variability was not shown to benefit generalisation for either children or adults in a phonetic training task with L2 vowels. This was surprising given the theory that being exposed to variability across voices should help generalise at least across voices (if not also across items), and that an HV benefit had been attested in the literature. Although some of the tasks showed little to no improvement from pre-test to post-test making it harder to draw solid conclusions, learning was seen in some contexts, and it is not clear why no variability effect was found there. However, the previous chapter discussed a few potential factors that could have played a role in obscuring a possible variability effect. One evident factor is the use of a single training session rather than the more usual multiple sessions. This could have affected the amount of learning that took place at all, as well as the participants' ability to adjust existing phoneme categories. The one-session paradigm also did not allow for any sleep consolidation, and limited the amount of time children had the chance to learn in the task after adjusting to the task demands. Another point that followed from Chapter 3 was that further research is needed to establish what type of learning materials and training paradigm might best facilitate children's learning. The use of meaningful pictures proved successful, but there could be a potential benefit in

also including *familiar* orthography in combination with those pictures. As discussed in Section 1.3.1, adding orthography to a task in which learners acquire a non-native phoneme contrast has been seen to be potentially helpful (Escudero et al., 2008). Moreover, for adult L2 learners, adding orthography to a picture-naming task has been found to be advantageous for their L2 vowel production (Nimz & Khattab, 2019), and orthographic information consistent with the phonological form has been found to be helpful in learning novel L2 words (Hayes-Harb, Nicol, & Barker, 2010). However, at the same time (Escudero, 2015; Simon et al., 2010) showed that for adult L2 learners, having orthography in an L2 speech training task did not seem to be helpful. The role of orthography has not been directly looked at with children learning L2 phoneme contrasts, but there is evidence that orthography is helpful in an L1 context. In learning novel words in their L1, 8-year-old children have been shown to benefit from having orthography present when being trained on the meaning of pictures, even if they are never told to make use of the orthography (Ricketts, Bishop, & Nation, 2009). Given that the paradigm using meaningful pictures is akin to word learning, it would be interesting to see if this benefit of orthography presence found for children learning novel L1 words also extends to L2 learning. Therefore, the current study aimed to extend the training from one to multiple sessions, include both pictures and orthography in learning, and use a more extensive pre/post-test battery to tease apart the factors that contributed to learning. The key question still remains

whether children show the same HV advantage as adults have been thought to do when using the HVPT paradigm for L2 speech sound learning.

The current study, rather than working with native English adults and children, moves to work with native Dutch primary school children of two ages, 7-8 year-olds and 11-12 year-olds, teaching them English vowel contrasts using a two-week training paradigm. The choice to work with Dutch learners of English had two motivations: first, Dutch schools are highly motivated with respect to English learning and it was anticipated that this would be necessary in order to recruit schools for such a lengthy study. Second, Dutch children are expected to have some existing knowledge of English from school, including some knowledge of the orthography. With regard to the choices of age-group, while initially, the plan was to test 7-8 year-olds and adults in line with the previous chapter, a pilot study showed even 16-year-olds had too much knowledge of English to be able to use the same task and stimuli as for the 7-8 year-olds, so 11-12 year-olds were chosen instead. Learners of this age are an interesting age group to investigate for two reasons. Firstly, in both the UK and the Netherlands age 11 seems to be a point at which L2 language teaching has long been implemented by the government: in the Netherlands children start learning an L2 at 10-11 while in the UK this used to be 11-12 (though note that this changed to 7 years in 2014). Secondly, 11-12 year-olds are at a pivotal age in terms of school phase: in the UK children have just started secondary school, and in the Netherlands they are in their final year of primary school preparing to move to secondary school, during which children in

both countries will receive more consistent L2 teaching. Participants were given a battery of pre/post-tests to investigate their discrimination and identification abilities, as well as their vocabulary learning and their production. The training paradigm was adapted from that developed in Study 1, using both pictures as well as orthography. As in the previous study, data was collected during the training task in order to track participants' performance across sessions.

In contrast to Study 1, in the current study participants were trained on multiple vowel contrasts, namely /ɛ/-/æ/, /ʌ/-/ɒ/, and /uː/-/ʊ/, which are known to be hard for Dutch learners of English (see Chapter 2 for a more detailed discussion). The reason for including multiple contrasts was that training on a full vowel set, which maps out the overall vowel space for the learner, has been shown to be beneficial over training on a subset or single contrasts in leading to greater generalisation (Nishi & Kewley-Port, 2007, 2008) as discussed in Section 1.3.2.

Acquisition of the phoneme contrasts was primarily investigated through means of a set of perceptual tasks. The use of a discrimination task is motivated, as in Study 1, following literature suggesting its efficacy in tapping into participants' ability to distinguish non-native phonemes (Logan & Pruitt, 1995), and also for the purposes of comparison with previous studies with children by Evans & Martín-Alvarez (2016), Giannakopoulou et al. (2017) and Study 1 in the current thesis. This study also included two sets of identification tests, which were conducted pre- and post-test (which was possible here given that children had some previous exposure to

English). Importantly, separate tests of orthography identification and picture identification were included, in order to help pull apart which of the two children might be relying on more during training. Finally, a category boundary task, which located the boundary at which learners' perception switched from one vowel to the other, was used to probe deeper changes in the location of the phonemes' best exemplars and their phoneme boundaries. As in Study 1, all of these pre-/post-tests used untrained talkers, meaning that improvement from pre- to post-test would indicate participants generalise to novel talkers.

In addition to these perception tasks, a production task was included in the pre-/post-tests in the form of real-word repetition. The production task was included to see whether perception training would also improve production for child participants. Recall from Section 1.3.3 that there have been various studies suggesting perceptual training might transfer to production for adults (Bradlow et al., 1997; Iverson et al., 2012; Shinohara & Iverson, 2018), but that few studies have tested this in children (Evans & Martín-Alvarez, 2016; Taimi, Jähi, Alku, & Peltola, 2014). Another reason for investigating production in this study is to see if variability training in particular has an effect on production. Recall that there has been little research into this, and what is out there is inconsistent (Brosseau-Lapré, Rvachew, Clayards, & Dickson, 2013; Evans & Martín-Alvarez, 2016; Kartushina & Martin, 2019). The current study adds to this small literature by investigating the role of perceptual variability training on production, using untrained talkers and items to test generalisation.

Finally, to investigate how well children are actually learning the meaning of the words as well as learning the individual vowels, two vocabulary tasks were included. Recall that the vocabulary learning literature predicts an HV advantage for vocabulary learning (see e.g. Barcroft & Sommers (2005); Rost & McMurray (2009); Sommers & Barcroft (2007, 2011)), This prompts further investigation as to whether variability encountered in phonetic training would transfer to vocabulary acquisition. The key vocabulary test in this study is the Vocabulary Task, where participants were played the stimuli with their respective pictures, and were asked for the meaning of the words (before being told the answer). At pre-test, this task also served the purpose of ensuring children understood the meaning of the words they were to be trained on. In addition, knowledge of vocabulary was also tested using the Picture Identification task, as in Experiment 2 of the previous study. This task also tests knowledge of the form-meaning association. However, for minimal pair test items, the task also probes discrimination ability. Therefore, as in Study 1, test items were included in this task in which the target and foil are *not* minimal pairs but instead differ in all phonemes, testing vocabulary learning. This is potentially easier than the Vocabulary task, which asks for Dutch translations of the words and may be harder for younger children in particular. As for the perception tasks, these two tasks used untrained talkers only to probe generalisation.

In addition to experimental tasks, each of the children also took a test of phonological working memory. This was used to check that the HV and LV groups

were balanced, and some exploratory analyses with this variable are reported. For this study, the hypotheses were pre-registered in advance at https://osf.io/kmspq. There are four key research questions which are asked for each of the tests of learning:

1. Does each age group improve during training / from pre-(training) test to post-(training) test? Improvement is predicted on all tasks, although differences in the extent of improvement in each task are expected.
2. Do older children and younger children differ in the amount that they improve during training / from pre-test to post-test in each test? Since evidence in the literature is mixed as to whether to expect younger learners to outperform older learners or vice versa, both the prediction that older children will show greater improvement than younger children (due to better ability to deal with the test tasks) and that younger children will show greater improvement than older children (due to plasticity, cf. Shinohara (2014)) will be tested.
3. For each age group, is there an effect of variability in the extent of improvement that is seen in training / from pre-test to post? For training, a low variability benefit is hypothesised, following Giannakopoulou et al. (2017). For the pre/post-tests, two hypotheses will be tested: both the hypothesis that the HV group show greater improvement with generalisation to novel speakers and items (following Barcroft & Sommers (2005); Lively et al. (1993)) and also the hypothesis that there is an LV

advantage (following the results with child learners in Giannakopoulou et al. (2017)).

4. Is the effect of variability different for younger children than it is for older children? This will specifically test for evidence that older children show more of an HV benefit, while younger children show no variability benefit or an LV benefit (following the age difference in Giannakopoulou et al. (2017); Shinohara (2014); Sinkevičiūtė et al. (2019)).

Note that the approach to hypotheses concerning age is a little different to that in Study 1, since specific directional hypotheses were pre-registered and tested. In addition to the above key hypotheses, for each dataset, overall performance for each age group was also tested against chance (except vocabulary and production tasks, where there is no "chance" level). Each analysis also compared whether participants differed at pre-test (with older children predicted to outperform younger children, due to greater pre-existing knowledge and/or better ability to deal with the tasks). Finally, although all of the test items in each task involve generalisation to an untrained speaker, *item novelty* was also manipulated for each of the tests where this is possible (i.e. for Discrimination, Orthography Identification and Production). The reason for including this manipulation was the same as in Study 1: although the variability over items is matched across conditions, it is possible that varying speaker-specific cues might also promote generalisation across this dimension (recall that evidence for this was seen in Evans & Martín-Alvarez (2016)). If this is the case, an interaction is expected such that

any high variability benefit would be stronger for untrained items than trained items. Therefore, this was also tested in all the tasks where item novelty was manipulated.

The study, including its hypotheses and the analysis plan, was preregistered https://osf.io/kmspq. Where analyses deviated from the pre-registered plan, this is noted (and justified) where the relevant analyses are reported.

4.2 Methods

All tasks used in this study were piloted for feasibility on two groups of Dutch children (distinct from those tested for the main study): a group of 13 11-12 year-olds (mean age = 11;11 years, SD = 5 months) as well as a group of 27 17-18 year-olds (mean age = 17;5 years, SD = 7 months). While both age groups could manage all of the tasks, it turned out that the 17-18 year-olds found the task too easy, leading to ceiling effects. Instead, as explained above, 7-8 year-olds and 11-12 year-olds were tested.

Although the preregistered experiment includes a Category Boundary task intended to investigate the location and slope of participants' category boundaries

for the DRESS-TRAP and STRUT-LOT vowel contrasts[14], this task malfunctioned during testing and could not be used to collect data. For brevity and ease of understanding, this task is not further described in the methods section; details about the stimuli and procedure can instead be found in Appendix VII.

4.2.1 Participants

Participants were Dutch primary school children, recruited from a school in the south of the Netherlands. Children from two years were recruited: children from *Groep 4* (the Dutch equivalent of UK Year 3, where children are 7/8 years old), and *Groep 8* (the Dutch equivalent of UK Year 6, where children are 11/12 years old). The school Head Teacher consented to their school participating in the research, after which individual participants were recruited through means of opt-out parental consent. A child could decide not to participate in a session at any time, and parents/carers could withdraw their child from participation at any point as well.

The study took roughly two weeks in total for each participant (12 sessions over a maximum of 12 consecutive school days, with some children receiving both pre-

[14] Note that the GOOSE-FOOT contrast was not used in this task because while performance was fine in an adult pilot, piloting the task with Dutch 11-12 year-olds and 16-17 year-olds revealed that a large portion of the synthetic vowel continuum was perceived as a different vowel altogether (/ə/), meaning children were not able to converge on a category boundary. FLEECE-THOUGHT was not used as this control contrast should not lead to categorisation difficulties.

test or post-test sessions in one day rather than two). In return for participation, children received a sticker at the end of each session, which they collected on a certificate they got to take home at the end of the study. The participating school received a €25 book voucher for their participation.

Our preregistered target sample size was 48 7-8 year old participants and 48 11-12 year old participants, as counterbalancing required us to test participants in multiples of 24 per variability condition. However, as noted in the preregistration, the practicalities of testing in school might lead to us testing more than 48 participants for each age group if a school was recruited in which all children in the class wished to take part (since children should not feel left out). This in fact occurred, and a total of 109 children was tested: 51 from *Groep 8* (29 HV, 29 LV) and 58 from *Groep 4* (25 HV, 26 LV).

In line with the preregistration, from this dataset a further 21 participants were excluded from the final sample: 1 due to not completing at least 60% of training (1LV 11-12 year-old), 12 due to diagnoses of dyslexia, hearing impairments, or other language disorders[15] (1 HV and 2 LV 7-8 year-olds, 4 HV and 5 LV 11-12 year-olds), 8 due to having English-speaking parents/carers or having lived in an

[15] Note that 7-8 year-olds were indicated as likely having DLD/dyslexia by their class teacher rather than having an official diagnosis, as children can only start the diagnostic process from age 7 onwards in the Netherlands.

English-speaking country for more than two months[16] (1 HV and 4 LV 7-8 year-olds, 1 HV and 2 LV 11-12 year-olds; stays abroad ranged from 1-5 years; 7 children had lived in the US, 1 in Hong Kong). This left a sample of 89 children. Of those, 50 were tested from *Groep 4* (mean age = 6;4, SD = 4 months, range 7;10-8;3 years) and 39 were tested from *Groep 8* (mean age = 6;6, SD = 4 months, range = 10;9-12;0 years).

In addition to these general exclusion criteria, for pre/post-tasks participants were excluded on a task-by-task basis as preregistered: participants were only included if their data was collected for both the pre-test and the post-test for that task. Note that one entire *Groep 4* class could not be administered the production post-test and phonological working memory task due to testing time restrictions in the class schedule. Additionally, one of the testing laptops used for recording the production (pre- and post-test) and phonological working memory (post-test only) tasks malfunctioned and did not save the files properly, resulting in dropouts in both tasks (13 *Groep 4* and 16 *Groep 8* for production, 7 *Groep 4* en 8 *Groep 8* for phonWM). For Production, if a participant's files were not saved for pre-test or

[16] For children who spoke more languages than just Dutch at home, I checked for each language whether it used the vowel contrasts that participants were trained on in English. These languages were Italian (1 7-8 year-old), Mandarin Chinese (1 7-8 year-old), Polish (1 11-12 year-old, 1 7-8 year-old), Russian (1 11-12 year-old), Thai (1 7-8 year-old), and Turkish (4 11-12 year-olds; 2 7-8 year-olds). None of these languages used phonemes of the contrasts of interest beyond those present in Dutch, so we decided to not exclude these children: all have unfronted /u/ and /i/, Turkish, Italian and Russian have /e/ while Thai and Polish have /ɛ/.

post-test, the entire set had to be excluded. The exact number of participants that was included for the analysis of each task is indicated in Table 43 below[17].

	Training	DM	vocab intro	picID	orthID	production	phonWM
All	89	82	81	80	80	33	42
G4	50	46	45	45	45	9	15
HV	27	26	25	25	25	6	9
LV	23	20	20	20	20	3	6
G8	39	36	36	35	35	18	27
HV	20	20	20	20	20	8	15
LV	19	16	16	15	15	10	12

Table 43. Participant numbers included in the analyses for training and each of the pre/post-tasks, split by age group and training variability condition.

The two age groups differed in the amount of exposure to English they had had in school: 7-8 year-olds had not received any English lessons yet, while 11-12 year-olds had received a one-hour-long English lesson per week for about a year. Additionally, children were asked about their English usage (see Appendix IV), for which the results can be seen in Table 44. It can be seen that that while 11-12 year-

[17] Ideally, participants who were excluded at the analysis stage for the reasons specified above would have been replaced to ensure equal numbers of participants in each of the experimental versions. However, for practical reasons this was not possible (i.e. school schedules, number of pupils available in the school, testing personnel availability, and the need for a continuous two-week testing period without inset days).

olds had more general exposure to English than 7-8 year-olds, the variability conditions within the groups were matched.

English use	Never	Hardly	Sometimes	Often	Always
7-8 year-olds					
TV/film	5 HV | 5 LV	6 HV | 6 LV	7 HV | 5 LV	5 HV | 5 LV	3 HV | 2 LV
Computer	5 HV | 7 LV	4 HV | 4 LV	4 HV | 1 LV	11 HV | 10 LV	2 HV | 1 LV
Music	5 HV | 4 LV	0 HV | 1 LV	3 HV | 3 LV	8 HV | 6 LV	10 HV | 9 LV
Reading	15 HV | 13 LV	4 HV | 3 LV	5 HV | 6 LV	2 HV | 0 LV	0 HV | 1 LV
11-12 year-olds					
TV/film	0 HV | 1 LV	5 HV | 0 LV	3 HV | 3 LV	10 HV | 12 LV	2 HV | 3 LV
Computer	2 HV | 2 LV	2 HV | 2 LV	3 HV | 5 LV	5 HV | 6 LV	8 HV | 4 LV
Music	0 HV | 2 LV2	0 HV | 2 LV	1 HV | 0 LV	3 HV | 3 LV	16 HV | 14 LV
Reading	11 HV | 5 LV	4 HV | 10 LV	4 HV | 3 LV	1 HV | 1 LV	0 HV | 0 LV

Table 44. Children's self-reported exposure to English in various situations.

4.2.2 Stimuli

Stimuli recordings were made in a sound-attenuated booth at UCL at a sampling rate of 44 100 16-bit samples per second. They were later downsampled to 22 050 samples per second, had their intensity scaled to 70 Hz, and were filtered with a band-pass filter using Praat (Boersma & Weenink, 2015). Stimuli were presented to speakers using ProRec (Huckvale, 2016). All stimuli can be found on the OSF: https://osf.io/bgdxp/.

Trained stimuli

Trained stimuli consisted of 16 monosyllabic two-way minimal pairs, all real English CVC-words. The defining difference for each pair was in the vowel, with

the consonant context remaining the same within each pair. Consonants used in the minimal pairs were present in the consonant inventory of both Dutch and English (see Chapter 2 for a detailed phonetic description of Dutch and English). This was done to ensure the focus of the training would be on the key difference, namely the 'foreign' English vowels that are absent from the Dutch vowel inventory[18]. The minimal pairs were divided up into 4 sets of vowel contrasts: GOOSE-FOOT /uː/-/ʊ/, STRUT-LOT /ʌ/-/ɒ/, DRESS-TRAP /e/-/æ/, and the control contrast FLEECE-THOUGHT /iː/-/ɔː/ (see Chapter 2 for in-depth discussion). The control contrast was included as a way to check whether participants understood the task, and to make the overall task more appealing by introducing a contrast they would be able to distinguish clearly.

Each vowel contrast was exemplified by 4 minimal pair items (8 words), resulting in a total of 32 trained items, or 16 trained pairs. Words were chosen on the basis of imageability since all training items consisted of both the orthography and a matching clipart picture representing their meaning (see Appendix VI). All words were deemed appropriate for primary school children. Care was taken to choose

[18] Note that while all consonants were present in the Dutch inventory, their distribution in terms of phonetic context may differ slightly (e.g. there are voiced plosives in syllable-final position while they would not occur in this position in Dutch). Key here is that the consonants would not be parsed as foreign per se.

words whose orthography for the vowels was transparent, and as consistent as possible.

Vowel contrast	Trained real-word items	Novel real-word items for Discrimination	Novel pseudo-word items for Orthography Identification
DRESS-TRAP	bed[a]-bad[b]	beg-bag	zem-zam
	gem-jam[b]	pet[d]-pat	zeb-zab
	pen[ab]-pan[b]	hem[d]-ham[b]	shen-shan
	vet[d]-vat[b]	said-sad	sheb-shab
GOOSE-FOOT	fool-full	should-shooed	suke-sook
	Luke[e]-look[d]	poot-put	sool-sull
	pool[d] -pull	could-cooed	jool-jull
	suit-soot	would-wooed	vuke-vook
STRUT-LOT	bus[b]-boss	gun-gone	zup-zop
	cut[c]-cot	lull-loll	fut-fot
	luck-lock	shuck-shock[e]	fum-fom
	shut-shot	cup[e]-cop[d]	bul-boll
FLEECE-THOUGHT	heel[a]-hall[b]	lead-lord	neek-nawk
	sheet-short	peak-pork	jeet-jort
	week[b]-walk	feel-fall	peeb-porb
	wheel[a]-wall	dean-dawn	teeb-torb

Table 45. Stimuli items for the Training, Discrimination, and Orthography Identification tasks.
Cognate status: [a] Spoken cognate; [b] written cognate; [c] spoken 'false friend' (same pronunciation, different meaning); [d] written 'false friend' (same written word, different meaning); [e] loan words used in Dutch, note 'Luke' is similar to the Dutch name 'Luuk' or 'Luc' /lyk/, especially when the English vowel is fronted.

While the aim was to avoid the use of Dutch-English cognates, this turned out to be impossible when relying on imageability and choosing child-appropriate vocabulary. However, training stimuli were either written cognates with only the pronunciation differing (often crucially in the non-native vowel), or more rarely

spoken cognates only where the writing differed from Dutch. The latter only happened for the FLEECE and DRESS vowels, which are very close in vowel quality to the Dutch counterparts (as discussed in Chapter 2). Consequently, the production of these vowels would be target-like regardless, so the use of spoken cognates here was deemed acceptable. Table 45 above lists the final trained stimuli; cognates are indicated with superscripts. These stimuli were used in Training, as well as in the Vocabulary task, Discrimination task, Orthography Identification task, Picture Identification task, and Production task. In Training and in the Vocabulary task these trained items were used with their corresponding clipart and orthography on screen. For the Discrimination task, these items were only presented aurally. In the Orthography Identification task, the stimuli were only presented with their orthography but not clipart picture on screen, while for the Picture Identification the clipart but not the orthography was presented.

Vowel contrast	Novel real-word pair combinations	Vowel contrast	Novel real-word pair combinations
FLEECE-TRAP	sheet-vat	DRESS-FOOT	bed-full
GOOSE-THOUGHT	Luke-walk	THOUGHT-STRUT	wall-bus
FLEECE-LOT	heel-shot	FOOT-LOT	pull-boss
STRUT-FOOT	luck-soot	TRAP-GOOSE	bad-fool
DRESS-STRUT	pen-cut	DRESS-THOUGHT	gem-hall
TRAP-THOUGHT	pan-short	TRAP-LOT	jam-cot
FLEECE-FOOT	week-look	GOOSE-STRUT	pool-shut
GOOSE-LOT	suit-lock	FLEECE-DRESS	wheel-vet

Table 46. Novel stimuli pairs for the Picture Identification task, and the respective vowel contrast they are testing.

The Picture Identification task also tested vocabulary (see Section 4.2.4). Therefore, in addition to presenting the trained item pairs, trained items were also presented in 32 novel picture pair combinations (i.e. in non-minimal pairs, see Table 46).

Novel stimuli

In addition to the 32 trained items, several of the pre/post-test tasks also included novel items to test generalisation ability to items beyond those on which participants were trained.

Discrimination

In addition to the trained items, the Discrimination task also used 32 novel items (16 pairs) to test if participants could generally hear the difference between the target vowels (see Table 45). All items used were real English minimal pair CVC words differing only in the vowel. Novel items were created using the same consonants used in the trained items to match their frequency.

Orthography identification

As seen in Table 45, the Orthography Identification task used 32 novel items (16 pairs), all of which were English pseudowords: nonwords that adhere to the phonotactic and orthographic rules of the language they are used for (Keuleers & Brysbaert, 2010). All novel items were minimal pairs differing in the vowel only, and were used to test generalisation, as well as to see if participants made use of the on-screen orthography when learning. Since English orthography is known to be opaque, the novel pseudoword items were created to be as transparent and

consistent as possible. Vowel orthography in particular was identical to that used in the trained items, but to avoid creating real words, some of the consonants were replaced with consonants that were transparent in their English pronunciation while not being foreign to Dutch phonology. All but 2 items were pseudowords in Dutch as well as English: *bul* and *fut* are real words in Dutch, however they are very low frequency (*fut* SUBTLEX frequency[19]: 0.34/million, *bul* SUBTLEX frequency: 0.39/million; Keuleers, Brysbaert, & New (2010)), and were therefore not expected to be familiar to our child participants.

Production – real-word repetition

Novel stimuli used for the Production task were 32 CVC-words with the same critical vowels as before, all of which were real English words. The novel items did not contain any consonant clusters to avoid unwanted transfer from Dutch (as described in Chapter 2, the variety of Dutch used by the participants has vowel epenthesis in certain consonant clusters). Only consonants that occurred in the set of trained items were used, again to match the consonant frequency. Four novel items were created per target vowel (see Table 47).

[19] For reference: the word with the highest frequency in SUBTLEX-NL is *ik* ('I'), at 39883.03/million, while the lowest frequency word is *1%te* at 0.02/million.

	TRAP	DRESS	GOOSE	FOOT	LOT	STRUT	FLEECE	THOUGHT
Novel real words for production	back	bell	boot	bush	toss	duck	bean	talk
	fat	wet	food	nook	not	mud	feel	call
	bat	neck	moon	wood	knock	sun	leak	fought
	tap	mess	tool	bull	pod	fuss	seat	lawn

Table 47. Novel real-word stimuli for the Production task, sorted by target vowel.

Phonological Working Memory task

Stimuli in the Phonological Working Memory task were replicated from the Dutch non-word repetition task produced by de Jong & van der Leij (1999; Scheltinga (1998), as based on the English Children's Test of Nonword Repetition (Gathercole, Willis, Baddeley, & Emslie, 1994). The Dutch adaptation of the test has 48 items and 2 practice items; items vary in length between 2 and 5 syllables (See Table 48).

For all items, primary stress was placed on the penultimate syllable, and secondary stress (present in five-syllable words only) fell on the first syllable. All vowels were short vowels, and to this end syllables were closed CVC combinations, as short vowels cannot occur in open syllables in Dutch.

Non-word item	Syllables	IPA
Practice: vepsim	2	ˈvɛpsɪm
2 syllables		
bippek	2	ˈbɪpɛk
detbik	2	dɛtbɪk
fusseg	2	ˈfʏsɛx
gannit	2	ˈɣɑnɪt
humdos	2	ˈɦʏmdɔs
jalvep	2	ˈjɑlvɛp
memmun	2	ˈmɛmʏn
mitsor	2	ˈmɪtsɔʁ
mukkef	2	ˈmʏkɛf
nembum	2	ˈnɛmbʏm
noggap	2	ˈnɔɣɑp
sigzef	2	ˈsɪxsɛf
3 syllables		
gikkallom	3	ɣɪˈkɑlɔm
harlonwig	3	ɦɑʁˈlɔnwɪx
kummigar	3	kʏˈmɪɣɑʁ
lemrospag	3	lemˈʁɔspɑx
nimmunnaf	3	nɪˈmʏnɑf
nomlunfam	3	nɔmˈlʏnfɑm
pigdulmek	3	pɪxˈdʏlmɛk
pippokket	3	pɪˈpɔkɛt
pirdapket	3	pɪʁˈdɑpkɛt
possallin	3	pɔˈsɑlɪn
savveffus	3	sɑˈvɛfʏs
vesrofsif	3	vɛsˈʁɔfsɪf

Non-word item	Syllables	IPA
Practice: meggof	2	ˈmɛɣɔf
4 syllables		
beffottirral	4	ˌbɛfɔˈtɪʁɑl
hissorreffum	4	ˌɦɪsɔˈʁɛfʏm
jafkertumsil	4	ˌjɑfkɛʁˈtʏmsɪl
kegjolbifmas	4	ˌkɛxjɔlˈbɪfmɑs
kepdarbitpuk	4	ˌkɛpdɑʁˈbɪtpʏk
lummoggappes	4	ˌlʏmɔˈɣɑpɛs
mifnemlunzan	4	ˌmɪfnɛmˈlʏnzɑn
ninnellummar	4	ˌnɪnɛˈlʏmɑʁ
nossiggeffas	4	ˌnɔsɪˈɣɛfɑs
peddattittup	4	ˌpɛdɑˈtɪtʏp
sirpegwotnal	4	ˌsɪʁpɛxˈwɔtnɑl
zosgefzilvas	4	ˌzɔsɣɛfˈzɪlvɑs
5 syllables		
bikkottaddukkep	5	ˌbɪkɔtɑˈdʏkɛp
bognupsarliftek	5	ˌbɔxnʏpsɑʁˈlɪftɛk
gaggollissuggef	5	ˌɣɑɣɔlɪˈsʏɣɛf
gambiskeflunjor	5	ˌɣɑmbɪskɛfˈlʏnjɔʁ
hussallimmoggep	5	ˌɦʏsɑlɪˈmɔɣɛp
lurreffannippos	5	ˌlʏʁɛfɑˈnɪpɔs
munfomlinzembam	5	ˌmʏnfɔmlɪnˈzɛmbɑm
nammonniffunnem	5	ˌnɑmɔnɪˈfʏnɛm
piptafbipketduk	5	ˌpɪptɑfbɪpˈkɛtdʏk
sorrammikkettul	5	ˌsɔʁɑmɪˈkɛtʏl
vugzasgoflifzef	5	ˌvʏxsɑsɣɔfˈlɪfsɛf
wogsirtumjafkel	5	ˌwɔxsɪʁtʏmˈjɑfkɛl

Table 48. Non-word stimuli items used in the Phonological working memory task. Number of syllables and broad IPA transcription of the East Brabantian Dutch pronunciation are provided for each item.

4.2.3 Design

Each participant completed the full training study, which involved three stages: pre-test, training, and post-test. The pre-test consisted of a Discrimination task, a Vocabulary task, a Picture Identification task, an Orthography Identification task, and a Production task. Training consisted of 8 sessions, four blocks each, spread across 8 days. The post-test was identical to the pre-test, with the addition of a phonological working memory task. The key experimental condition was once more the amount of talker variability in the training input, with the HV and LV conditions only differing in training (with identical tests).

For each variability condition, speakers were counterbalanced across participants. In total, 8 speakers were used throughout the experiment: female talkers F1, F2, F3, F4, and F5, and male talkers M1, and M2 are all native speakers of Standard Southern British English, while GB is a female native speaker of East Brabantian Dutch. Talkers used in training were not used in pre/post-tests and vice versa.

In Training, participants heard four different speakers (2 male, 2 female) in the HV condition, while in the LV condition stimuli were spoken by a single talker (one of the four used in HV). The four speakers in HV (M1, M2, F4, F5) were blocked, as trial-by-trial speaker adaptation has been shown to be detrimental to performance (Dong et al., 2019; Perrachione et al., 2011), and speaker order was counterbalanced across experiment versions so that each speaker occurred in every block (see Table 49). To match the training task across conditions, LV training was

also organised in blocks, but participants heard the same speaker across all four blocks.

	V1	V2	V3	V4
HV	F4 F5 M1 M2	F5 M1 M2 F4	M1 M2 F4 F5	M2 F4 F5 M1
LV	F4	F5	M1	M2

Table 49. Counterbalanced versions of Training: participants receive either the HV or the LV variant of one of the 4 training versions. Each row in the HV section corresponds to a block, presented in order.

Pre- and post-tests were identical across conditions, and all stimuli involved novel speakers (F1, F2, F3) who had not been encountered during Training. These speakers were rotated between the Picture Identification, Orthography Identification, and Production task. One additional novel speaker (GB) was used for the Dutch Phonological Working Memory task for all versions of the experiment (see Table 50).

	V1	V2	V3
Discrimination	F1 & F2 & F3	F1 & F2 & F3	F1 & F2 & F3
Vocabulary introduction	F1	F2	F3
Picture identification	F2	F3	F1
Orthography identification	F3	F1	F2
Production	F1	F2	F3
Phonological Working Memory	GB	GB	GB

Table 50. Counterbalanced versions of the pre/post-tests: participants are assigned one of the 3 versions. Note that the Discrimination task is identical across versions, as it uses all three novel speakers in each trial.

Table 51 describes which versions of training and pre-tests were combined for the 24 different counterbalanced versions (12 HV and 12 LV): each version is a combination of one of 8 versions of training (4 HV and 4 LV) and one of 3 versions of the pre/post-tests.

	Training	V1	V2	V3	V4
Pre/post					
V1		HV1 / LV1	HV4 / LV4	HV7 / LV7	HV10 / LV10
V2		HV2 / LV2	HV5 / LV5	HV8 / LV8	HV11 / LV11
V3		HV3 / LV3	HV6 / LV6	HV9 / LV9	HV12 / LV12

Table 51. Counterbalanced versions of the experiment showing which combinations of Training and pre/post-test versions they entail.

4.2.4 Procedure

The training and the perception tasks in the pre/post-test were run in Gorilla (www.gorilla.sc, Anwyl-Irvine et al. (2019)). These tasks were run on 30 Dell Vostro 15 3000 series laptops available in the school, and each child had their own in-ear headphones. The production task and phonological working memory task were run on 4 additional laptops, and were run using PsychoPy (Peirce, 2007), where children were provided with Sennheizer HD201 over-ear headphones. Testing took place in the children's own classroom or a relatively quiet room in the school. The researcher came to the school for three continuous weeks and administered all tasks with the children. Children first completed the pre-test, spread out over two sessions: the majority of tasks was completed in the classroom by the entire class at once, while the production task and questionnaire were administered in groups of 4 (during production, children were placed as far apart

as possible; the task randomised stimuli per participant so all 4 children heard a different stimulus at any given time). Training started on the same day for all classes, and lasted 8 sessions. Children received training in full class groups every school day where possible. The post-test took place the day after the last training session to ensure any temporary phoneme categories built up during training were no longer maintained (Heeren & Schouten, 2010). The post-test was again administered in two sessions with full class testing for the perception-based tasks and small group testing for the production task and phonological working memory task. See Figure 35 for an overview of the procedure.

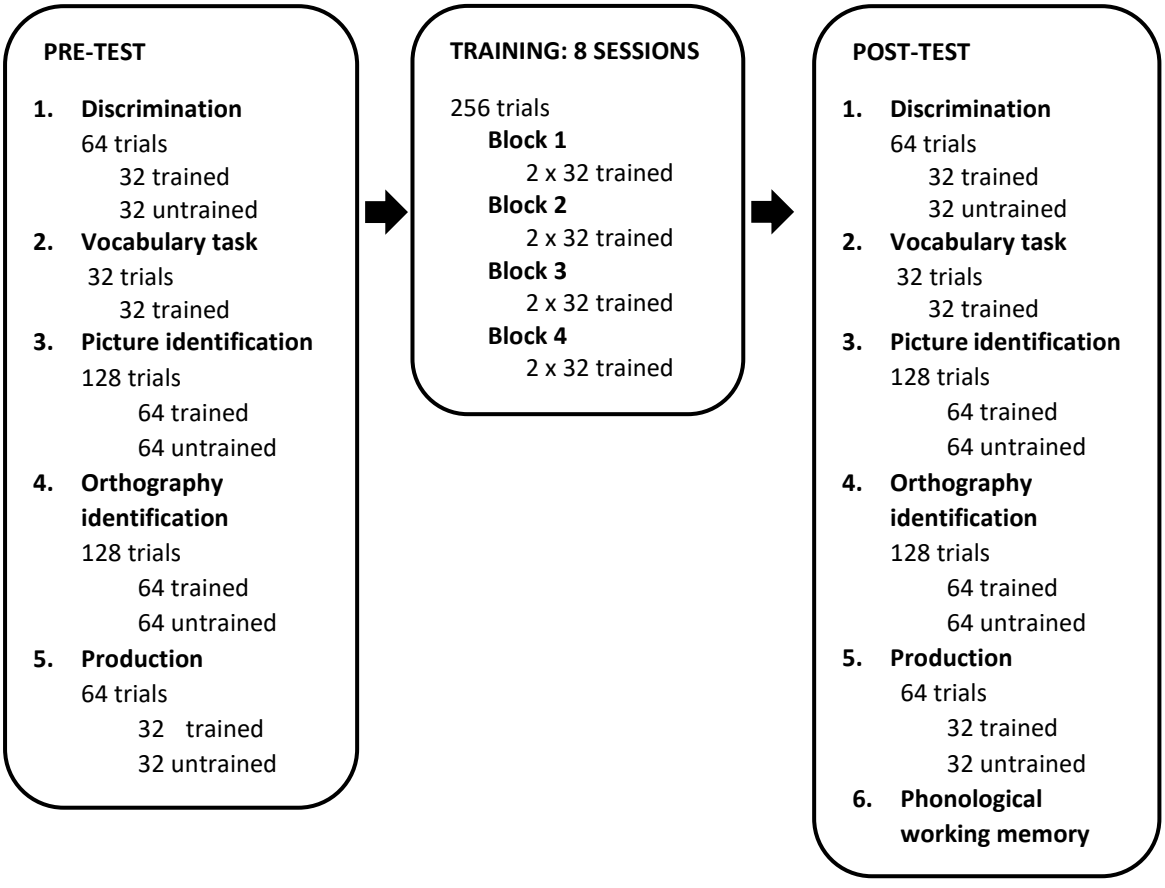


Figure 35. Overview of the procedure of Training and the pre/post-tests, including the number and type of experimental trials per task. Task numbering shows the order in which the tasks were conducted. Pre- and post-test were each administered over 2 sessions, with the Production and Phonological Working Memory tasks being administered separately from the other tasks.

Training

The training task was a 2AFC identification task. Training consisted of 8 sessions, of 4 blocks each. Each block comprised stimuli from a single talker (the talker changed for each block for the HV condition, but not for the LV condition; see Table 49 for counterbalancing), and comprised 64 trials – twice for each of the 32 training words as target. In total, 256 trials were presented per session. Trial order was randomised within blocks on a participant-by-participant basis. For the first session, participants were given a break between each block, but after the first session, the breaks between training blocks were taken out of the experimental task as they turned out to be distracting for the 7-8 year-olds and lead to some participants failing to complete the session.

In each trial, participants heard a word and selected one of two candidate options displayed on the screen (see Figure 36) by clicking the respective button. Participants were given trial-by-trial feedback on their performance[20]. If participants selected the correct answer, a green happy face was presented as well as the Dutch text *Hoera, je hebt een muntje verdiend!* ('Hooray, you won a coin!'). Simultaneously, the correct word was repeated and the correct option was shown. Additionally, the number of coins earned (indicated by a number in the bottom-

[20] After the first session, the time during which the feedback was on the screen was reduced from 1500 ms initially to 1100 ms instead, to slightly speed up the task after class feedback from the teacher.

right corner of the screen) was incremented. If an incorrect answer was selected, a red sad face was presented as well as the text *Jammer, volgende keer beter!* ('Too bad, better luck next time!'). Again, the correct word and option were repeated. The total number of coins a participant had won was always visible during the training trials (though not during feedback), as was the total number of trials. At the end of training, participants were shown a screen that indicated the total number of coins they had earned. The experimental program recorded participants' trial-by-trial accuracy.

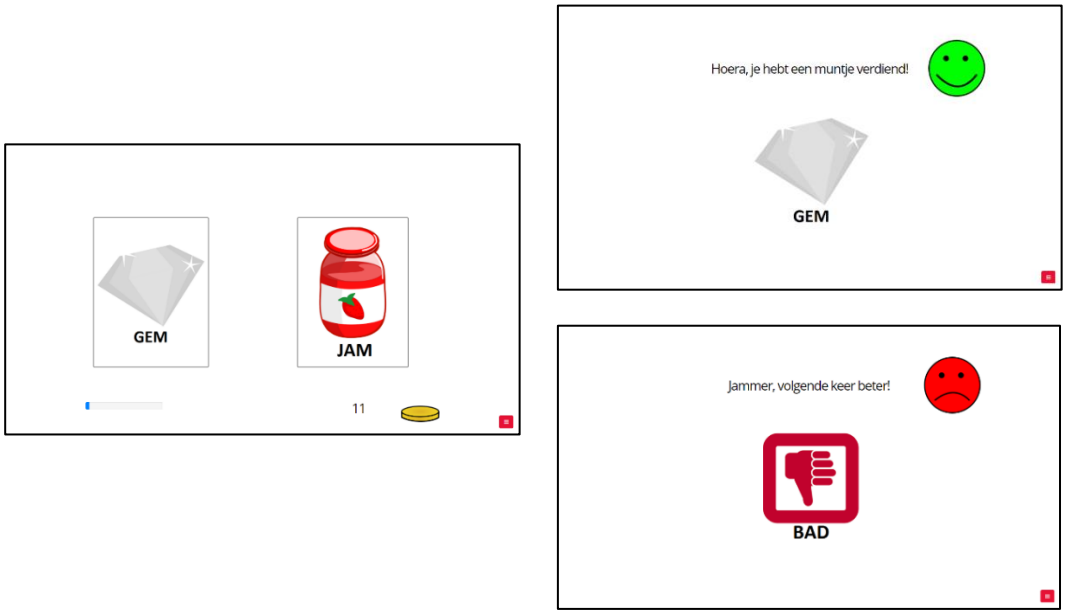


Figure 36. Trial representation for Training, with the top right screen showing the feedback after a correct response, while the bottom right screen shows the feedback after an incorrect response.

Pre/post-tests

Discrimination task

Discrimination ability was measured through means of a three-alternative forced-choice (3AFC) oddity task. In the task, three clipart pictures of a Queen's guard appeared on the screen (see Figure 37). Participants heard three words played (ISI

200ms), two of which were different tokens of the same word (e.g. '*gem*, '*gem*), and one of which was a different word but always part of a minimal pair with the other two (e.g. '*jam*). Participants indicated which of the three words was the odd one out by clicking on the corresponding picture. Each token was spoken by a different talker not used in training in order to test generalisation across talkers.

This task contained 16 pairs which also occurred in training (trained items) and 16 pairs which did not (untrained items) occurring twice each in random order, making 64 trials in total. Note that all test items probe generalisation to a novel talker, and the novel items tested generalisation abilities across items. Trial order was pseudo-randomised so that all vowels occurred equally as target and foils. The experimental program recorded participants' accuracy, and did not allow for a response before the third word had finished playing. No feedback was provided.

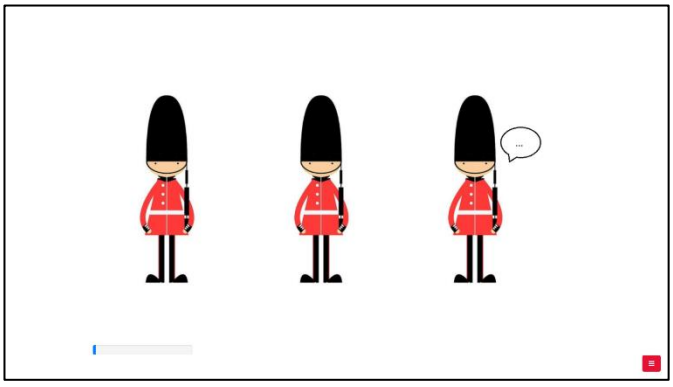


Figure 37. Trial example for the 3AFC Discrimination task.

Vocabulary task

In this task, participants heard the 32 individual items they were to be trained on once each in randomised order, spoken by a talker not used in training. For each trial (see Figure 38), participants heard the English word, saw the corresponding

clipart picture and English orthography, and were asked to type in the word's meaning if they could[21]. Then they were provided with the Dutch translation of the word, after which they were asked to provide a yes/no response as to whether they were familiar with the English word[22].

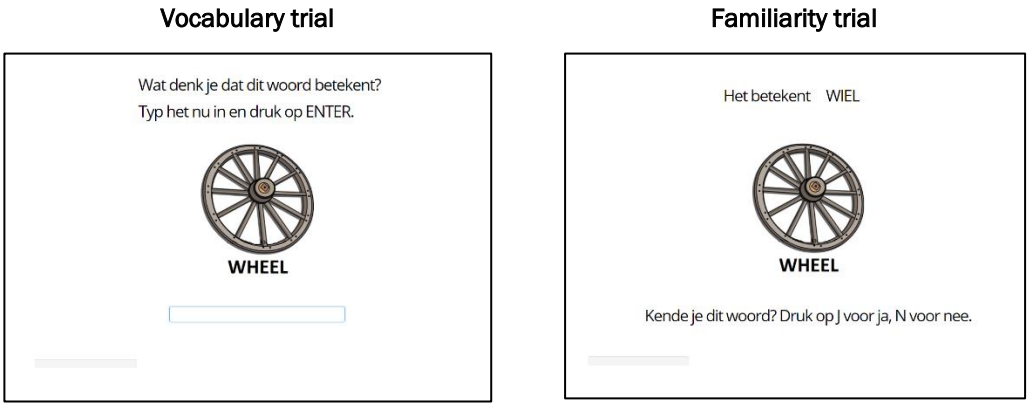


Figure 38. Trial example for the Vocabulary introduction task. Text on screen for Vocabulary trial: "What do you think this word means? Type it here and press ENTER". Text on screen for Familiarity trial: "This means [TRANSLATION]. Did you know this word? Press J for yes, N for no."

This task had two purposes. First, it was included at pre-test to familiarise participants with the words and to ensure they understood the meaning of each

[21] Note that this part of the task was particularly hard for the 7-8 year-olds, since they were still learning to write and spell in Dutch. In order to limit testing time, the 7-8 year-olds were instructed to only type in the word if they knew what it meant and knew how to write it, while the 11-12 year-olds were instructed to guess if they were not sure what the word meant.

[22] This familiarity question was intended to be used as a less complex way of measuring vocabulary knowledge that might be better suited for the 7-8 year-olds, leading to a 'familiarity score' for each trained item which could be used to explain variation in the models. However, this task seemed to confuse the children. Additionally, it could not be used to investigate novel words, making it inappropriate for most pre/post-test analyses. Its contribution to the models was tested for tasks involving only trained items (Vocabulary and Training, see the respective analysis scripts), but familiarity never improved the model fit. Therefore, the familiarity score will not be mentioned again in this chapter.

picture, as well as to measure their pre-existing vocabulary knowledge. Second, by including it at both pre- and post-test, it was possible to use this as a measure of vocabulary learning. Translation accuracy was determined by whether participants entered a reasonable translation of the target word, regardless of whether the translation was spelled correctly. Trial order was randomised, and no feedback was provided.

Picture identification task

In the Picture Identification task, a 2AFC identification task, participants were presented with two clipart pictures on the screen while hearing one spoken word corresponding to one of the two options (see Figure 39). Participants were asked to click on the picture corresponding to the word they heard, but were not given feedback. The task consisted of 128 trials: 64 minimal pairs used in training, as well as 64 non-minimal pair combinations items that differed in more than just the vowel, allowing these items to test vocabulary knowledge without requiring an ability to discriminate the vowels. Presentation of minimal pairs and non-minimal pairs was blocked, with trials presented in a randomised order within each section. All items were spoken by a novel speaker, and were repeated once to make both of the options of a pair the target. No feedback was provided. The experimental program recorded response accuracy.

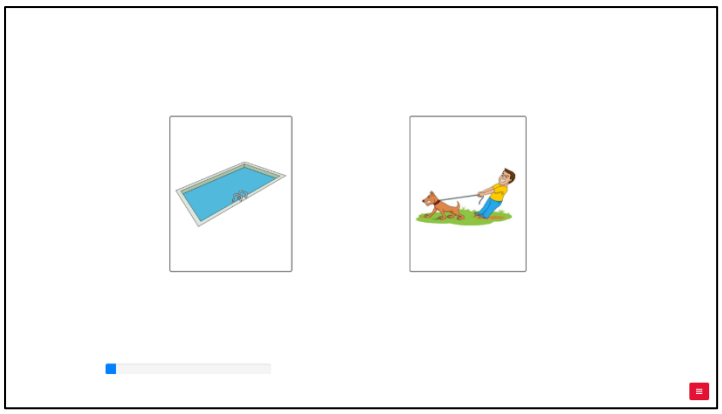


Figure 39. Trial example for Picture Identification. The left-hand picture depicts the word ‘pool’, testing the GOOSE vowel, while the right-hand picture depicts ‘pull’ for the FOOT vowel.

Orthography identification task

During the Orthography identification task, a 2AFC identification task, participants were presented with two written minimal pair words on the screen while hearing one spoken word corresponding to one of the two options (see Figure 40). Participants were asked to click on the word they heard, but were not given feedback. The task consisted of 128 trials: 64 items used in training, as well as 64 novel items with the same vowels to test generalisation. Presentation of trained and novel items was blocked, with trials presented in a randomised order within each section. All items were spoken by a novel speaker, and were repeated once to make both of the options of a pair the target. No feedback was provided. The experimental program recorded response accuracy.

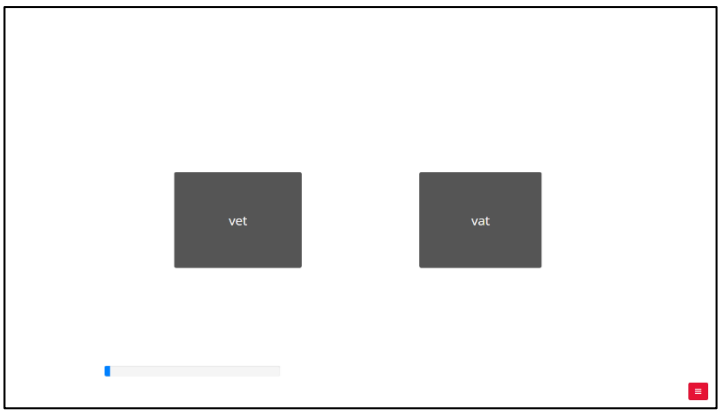


Figure 40. Trial example for Orthography Identification, displaying 'vet' versus 'vat' to test the DRESS-TRAP contrast.

Production - real-word repetition task

For the Production task, participants heard an English word, and were asked to repeat it out loud. No visual information was provided during the trial, apart from the trial number and total number of trials to enable participants to track their progress. The task included 32 trained items, as well as 32 novel items. Note that all test items probe generalization to a novel talker, and the novel items tested generalisation abilities across items. Trial order was randomised, and no feedback was provided. Participants' production accuracy was rated by a native speaker of SSBE (see Section 4.3.6 in the results for coding details).

Phonological working memory task (post-test only)

The Phonological Working Memory task was a computerised version of a standardised Dutch non-word repetition task as described in Scheltinga (1998): non-words were heard twice, after which the participant was asked to repeat them as correctly as possible. The task consisted of 2 practice trials, and 48 experimental trials presented in random order. No visual information was provided on screen,

apart from the trial number and total number of trials to enable participants to track their progress. Trial order was randomised, and no feedback was provided. An item is scored as correct if all syllables are correctly repeated.

4.2.5 Analyses and statistical approach

Data were analysed using both Frequentist and Bayesian statistics, and were preregistered. The preregistration document can be found here: https://doi.org/10.17605/OSF.IO/KMSPQ

Model building

Data were analysed using generalised logistic mixed effects models (Baayen et al., 2008) for accuracy data, and linear mixed effects models for some of the production data, using the *lme4* package (Bates et al., 2015) for the R environment (R Core Team, 2018). The *plyr* package (Wickham, 2011), *knitr* (Xie, 2018) and *kableExtra* packages (Zhu, 2018) were used to present the results. Graphical representations were created using *ggplot2* (Wickham, 2016) and *ggpirate* (Braginsky, 2018).

The approach to model building was, as in Study 1, 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 all tasks where this was manipulated, i.e. *trained/novel* for Discrimination, Orthography Identification and Production, *minimal pair/non-pair* for Picture Identification) and all the interactions between them. Following

the preregistration and the procedure in Study 1, the age groups were first analysed separately, and only after that were performances compared across the age groups, meaning that *age group* was only a relevant variable for the latter analysis. The approach to random effects was as in Study 1, with *participant* automatically included as a random effect with a full random slope structure – or the most complex to converge – while for items, by-item slopes were only included if they contributed to the model. For consistency, the final model structure detailing the random effects structure and included control variables is reported for each model.

As before, a centred coding was used for predicting variables (see the Analyses sub-section in Section 3.2.1). For factors with more than two levels which are not naturally ordered (i.e. *talker* and *vowel type*), a centred contrast coding was used – i.e. dummy variables were created for each of the contrasts compared with a reference level, but a centred coding was used, allowing the intercept and other effects to be evaluated as averaged over all levels of this variable, rather than at the reference level. To assess whether performance was above chance, the intercept was compared to chance level. All reported models converged using the Bound Optimization by Quadratic Approximation (BOBYQA optimization, Powell (2009)).

A difference between these analyses and those in the previous study is that, in addition to the experimental factors, factors to control for *Vowel contrast* and *Talker* were included in the analyses as these factors likely contribute variance to the data. It is clear that the particular vowel contrasts used in this study (DRESS-

TRAP, GOOSE-FOOT, STRUT-LOT, and FLEECE-THOUGHT) are treated differently perceptually by Dutch learners of English (see Chapter 2). Therefore, *Vowel contrast* is automatically included in the models as a fixed effect, though it is not included in any interactions. It is coded as a set of 3 centred contrasts (so that the intercept continues to represent the grand mean). As per the preregistration, 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 (as it essentially functioned as a control contrast, see Stimuli). If performance had been at ceiling, the data from that particular contrast would have been removed for the analyses of that task; this turned out to never be the case. Performance plots split by vowel contrast can be found in the Appendix VIII and can be found in the analysis script at: https://osf.io/bgdxp/. As per the pre-registration, inferential statistics regarding vowel-contrast are not reported or computed. However, the data have been made available online (https://osf.io/bgdxp/) allowing for future exploratory analyses concerning the different contrasts.

As described above, research suggests that *Talker* may potentially contribute variance to the model as well, depending on the idiosyncratic properties of the sample of talkers used in this experiment. Conceptually, *Talker* is more like a random effect, but these cannot have less than 6 levels (Bolker, 2019) so it is entered as a fixed effect here. Therefore, for the analyses here *Talker* was not automatically included but instead only included when it contributed to the model

fit (using a cut off of $p > .2$ rather than $p > .05$ similar to what is suggested for random effects by Matuschek et al. (2017)). Where it did significantly contribute to the model, *Talker* was entered as a fixed effect only; it was not interpreted and not included in interactions. *Talker* was coded as a set of 3 centred contrasts in the Training models (where the variable has 4 levels), and centred for all other fixed effects in the pre/post-test analyses (where it has 2 levels).

As part of the post-test task battery participants' *Phonological Working Memory* score was collected to account for some individual variation in overall test performance; a higher score indicates participants had a better phonological working memory. However, due to a large amount of missing data this score was not included in the models as was planned in the preregistration. Instead, exploratory correlational analyses were performed.

Inference criteria - Frequentist and Bayesian analyses

As for the study reported in Chapter 3, although models included all of the experimentally manipulated variables and their interactions, statistics are interpreted and reported only for the fixed effects relevant to the hypotheses (as laid out in Section 4.1 above).

For the frequentist analyses, all main effects and interactions in the models were interpreted as significant at an alpha level of .05. Alpha values for logistic mixed effects models were computed automatically by the *lme4* package (Bates et al.,

2015), while alpha values for linear mixed effects models were computed using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017).

As in Study 1, the key inferential statistic is the Bayes Factor (BF), computed using the method advocated by Dienes (2008, 2015) (see the Analyses sub-section in Section 3.2.1 for a detailed description). H1 is again modelled as a half normal in each case, with the standard deviation set to an estimation of the predicted effect size *x.* The key issue is once more how to determine appropriate values of x. In some cases, this was done by using estimates taken from equivalent analyses in Giannakopoulou et al. (2017), who used a similar task design. Where this was not possible or appropriate, as in Study 1, values from within the current data set were used. Wherever possible, these values were taken from independent data, by using the value for the equivalent effect for 7-8 year-old children to inform that for 11-12 year-old children and vice versa. In some cases standard deviations were computed using knowledge of constraints on the likely maximum value, such as a maximum of the scale. Some of these calculations require an estimation of minimal performance. For most tests this was chance, however, for tests with no notion of chance (i.e. in Vocabulary and Production), instead the value of getting one correct answer was used as the minimum (since the logodds/BF of 0 cannot be computed). This approach for estimating H1 was laid out in the preregistration. For the most part, this pre-registration was adhered to. However, there are places where, in hindsight, the pre-registered plan was not appropriate (either due to unexpected aspects of the data, or because there was an error in the original plan). Wherever

estimates deviated from the preregistration, this is transparently indicated and justified. The computation of the Bayes Factors reported in the tables can be found in Appendix III. Where BF results were reported in-text rather than in a results table, the notation $BF_{H(0,x)}$ is again used to denote a Bayes Factor where x is the sd of the half normal used to model H1.

As in Study 1, since there is subjectivity in the choice of value to inform H1, robustness regions are once more included. To compute the ranges, values were tested – in increments of 0.01 - from a difference of 0 from chance to the log-odds score corresponding to the difference between chance and ceiling performance (deemed to be all but one trial correct: i.e. 99.6% (log-odds 5.5156) in Training; 98.4% (log-odds 4.8299) in Discrimination; 99.2% (log-odds 4.84) in Orthography Identification; 98.4% (log-odds 4.1425) in Picture Identification; 96.8% (log-odds 6.8659) in Vocabulary). 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 (Discrimination), except for ranges of values giving evidence for H0, where the maximum is infinity. Robustness Regions should be interpreted bearing in mind larger values of H1 bias evidence for the *null,* whereas smaller values bias in favour of H1.

4.3 Results

Results will be discussed per task. For each task, the results are discussed per individual age group, before discussing the combined results looking at age differences. In the tables reporting the Bayes Factors, the justification for each of the predictors is indicated; these refer to the justifications in Appendix III. Colour coding has been used to help understand the results at a glance: green means the Bayes Factor shows evidence for H1, yellow means the Bayes Factor is ambiguous, and red means the Bayes Factor shows evidence for H0. Analyses reported here contain all vowel contrasts learners were trained on. Full analyses were also run without the FLEECE-THOUGHT control contrast in the data, but as the overall pattern of results remained the same these analyses can be found in Appendix IV.

4.3.1 Training

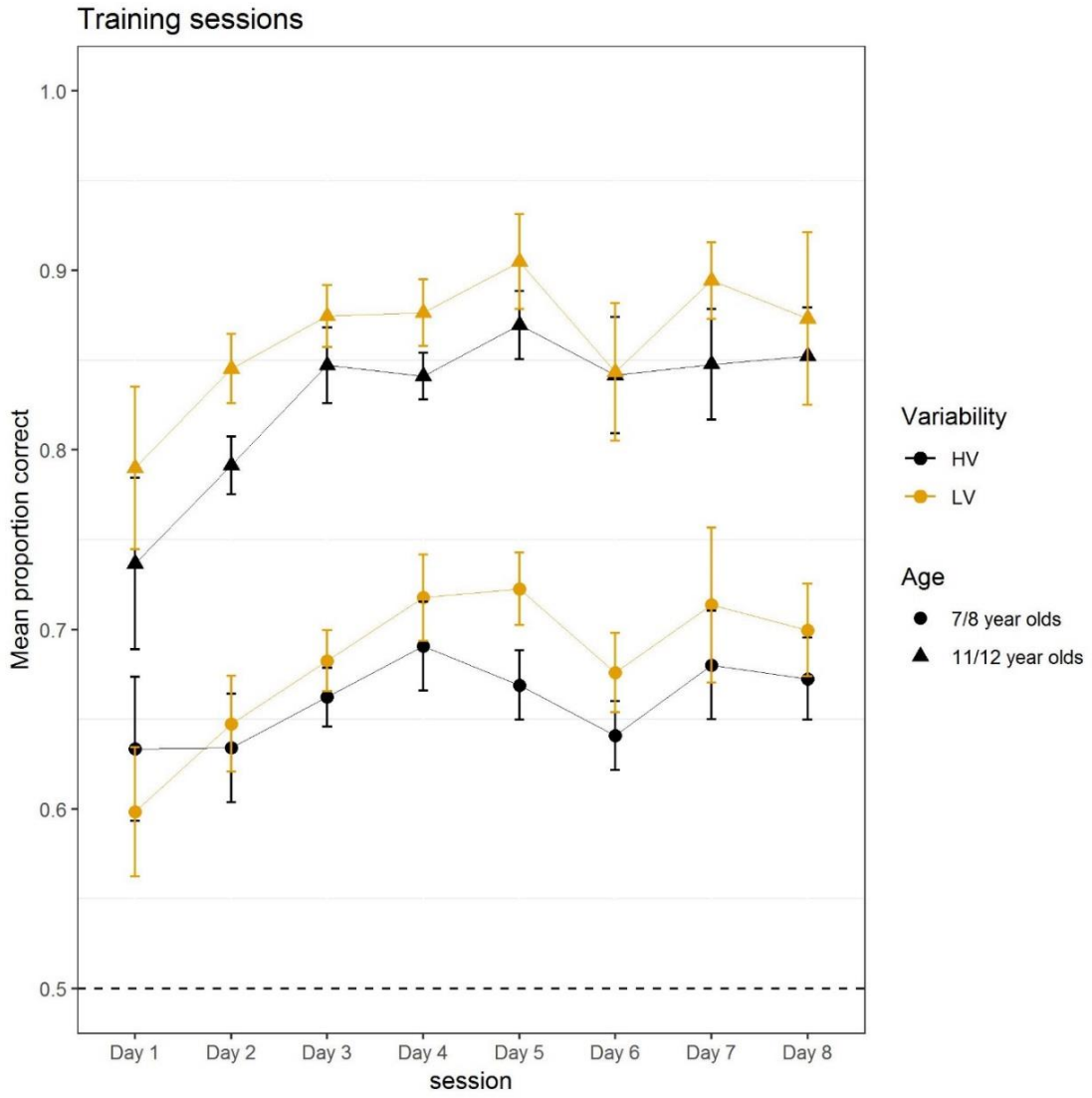


Figure 41. Accuracy results for 7-8 year olds and 11-12 year olds during Training of Experiment 5, comparing accuracy for HV versus LV training input. The error bars indicate 95% CI, and the dashed line indicates chance level.

7-8 year olds

As seen in Figure 41 and Table 52, there was substantial evidence that 7-8 year olds performed above chance in the Training task and that they improved across sessions (note that the dip in improvement at Day 6 occurred after a weekend). The evidence for LV outperforming HV overall and for there being more improvement in LV than in HV was ambiguous.

Final structure TrainG4 model: accuracy ~ session*condition + VowelContrast + (session:condition|participant) + (1|item)

Hypothesis	fixed effect in model	β	SE	z	P	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.806	0.089	9.076	<.001	2.362	$5.45*10^{16}$	[0, >5.5156]
improvement from pre- to post- test	session	0.033	0.010	3.370	.001	0.166	33.719	[0, 1.5156]
LV outperforming HV	condition	0.139	0.145	0.957	.339	0.648	0.557	[0, 1.1156]
greater overall improvement for LV training	Interaction condition: session	0.032	0.020	1.637	.102	0.103	1.3	[0, 0.3156]

Table 52. Mixed model results for the Training analysis, for 7-8 year olds.

11-12 year olds

As seen in Figure 41 and Table 53, there was substantial evidence that 11-12 year-olds performed above chance in the Training task, that they improved across sessions, and for LV outperforming HV overall. The evidence for there being more improvement in LV than in HV was ambiguous.

Final structure TrainG8 model: accuracy ~ session*condition + VowelContrast + talker + (session*condition|participant) + (session:condition||item)

Hypothesis	fixed effect in model	beta	SE	z	P	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	2.362	0.183	12.907	<.001	0.806	$1.11*10^{34}$	[0, >5.5156]
improvement from pre- to post- test	session	0.166	0.038	4.378	<.001	0.033	92.972	[0, >5.5156]
LV outperforming HV	condition	0.648	0.215	3.012	.003	1.626	22.596	[0.1156, >5.5156]
greater overall improvement for LV training	Interaction condition: session	0.039	0.076	0.513	.608	0.103	0.854	[0, 0.3156]

Table 53. Mixed model results for the Training analysis, for 11-12 year olds.

Differences between age groups

Comparing the two age groups (see Table 54), there was substantial evidence that 11-12 year-olds outperformed 7-8 year-olds across Training. In terms of the evidence for an effect of age on improvement across training sessions, there was substantial evidence for older learners outperforming younger learners and substantial evidence against younger learners outperforming older learners.

Final structure TrainAgeComp model: accuracy ~ session*condition*age group + VowelContrast + (session.ct:condition.ct*group.ct||participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Older outperform younger	Age group	1.370	0.160	8.537	<.001	1.998	$8.50*10^{14}$	[0.1156, >5.5156]
Older show greater improvement than younger	Interaction Session:age group	0.103	0.023	4.494	<.001	0.083	6328.34	[0, >5.5156]
Younger show greater improvement than older	Interaction Session:age group	-0.103	0.023	-4.494	<.001	0.083	0.045	[0, >5.5156]

Table 54. Mixed model results for the age comparison of the Training results.

4.3.2 Discrimination

7-8 year olds

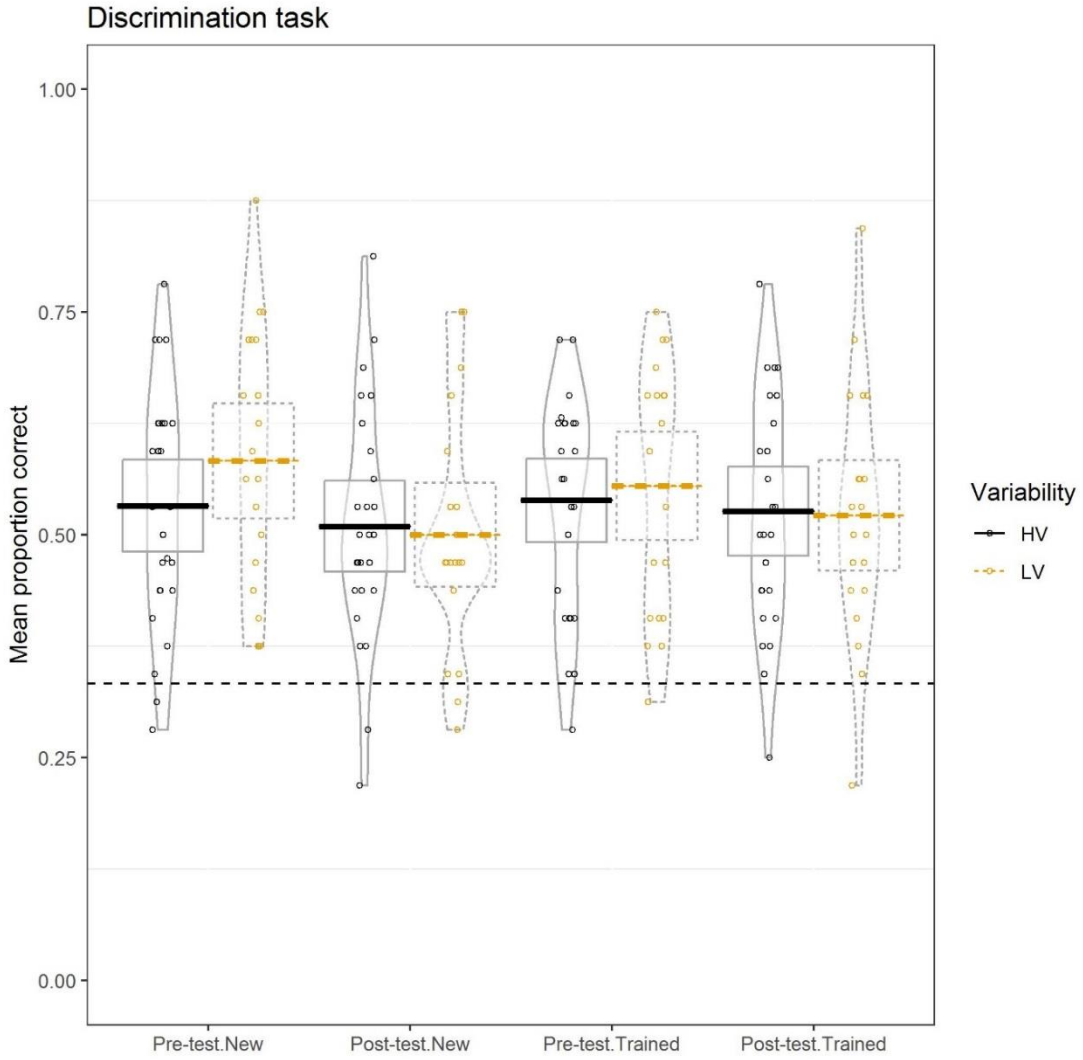


Figure 42. Accuracy results for 7-8 year olds on the pre- and post-test Discrimination task of Experiment 5, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 42 and Table 55, there was substantial evidence that 7-8 year olds performed above chance in the Discrimination task, but substantial evidence that they did not improve overall from pre- to post-test. The evidence for there being more improvement in HV than in LV was ambiguous, but there was substantial

evidence against there being more improvement in LV than in HV. There was ambiguous evidence for there being greater improvement for HV than LV for novel items, but no such difference for trained items.

Final structure DiscrimG4 model: accuracy ~ session*condition*itemnovelty + VowelContrast + (session:condition:itemnovelty|participant) + (itemnovelty|item)

Hypothesis	fixed effect in model	beta	SE	z	P	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.833	0.096	8.686	<.001	1.689	$2.43*10^{15}$	[0, >4.8299]
improvement from pre- to post- test	session	-0.154	0.055	-2.785	.005	0.219	0.063	[0.1299, >4.8299]
greater overall improvement for HV training	Interaction condition: session	0.190	0.112	1.704	.088	0.219	2.689	[0, 2.6299]
greater overall improvement for LV training	Interaction condition: session	-0.190	0.112	-1.704	.088	0.219	0.183	[0.1299, >4.8299]
greater overall improvement for HV training for novel items, no diff for trained	Interaction Condition: session: novelty	-0.173	0.224	-0.774	.439	0.146	0.648	[0, 0.3299]

Table 55. Mixed model results for the Discrimination task, for 7-8 year olds.

11-12 year olds

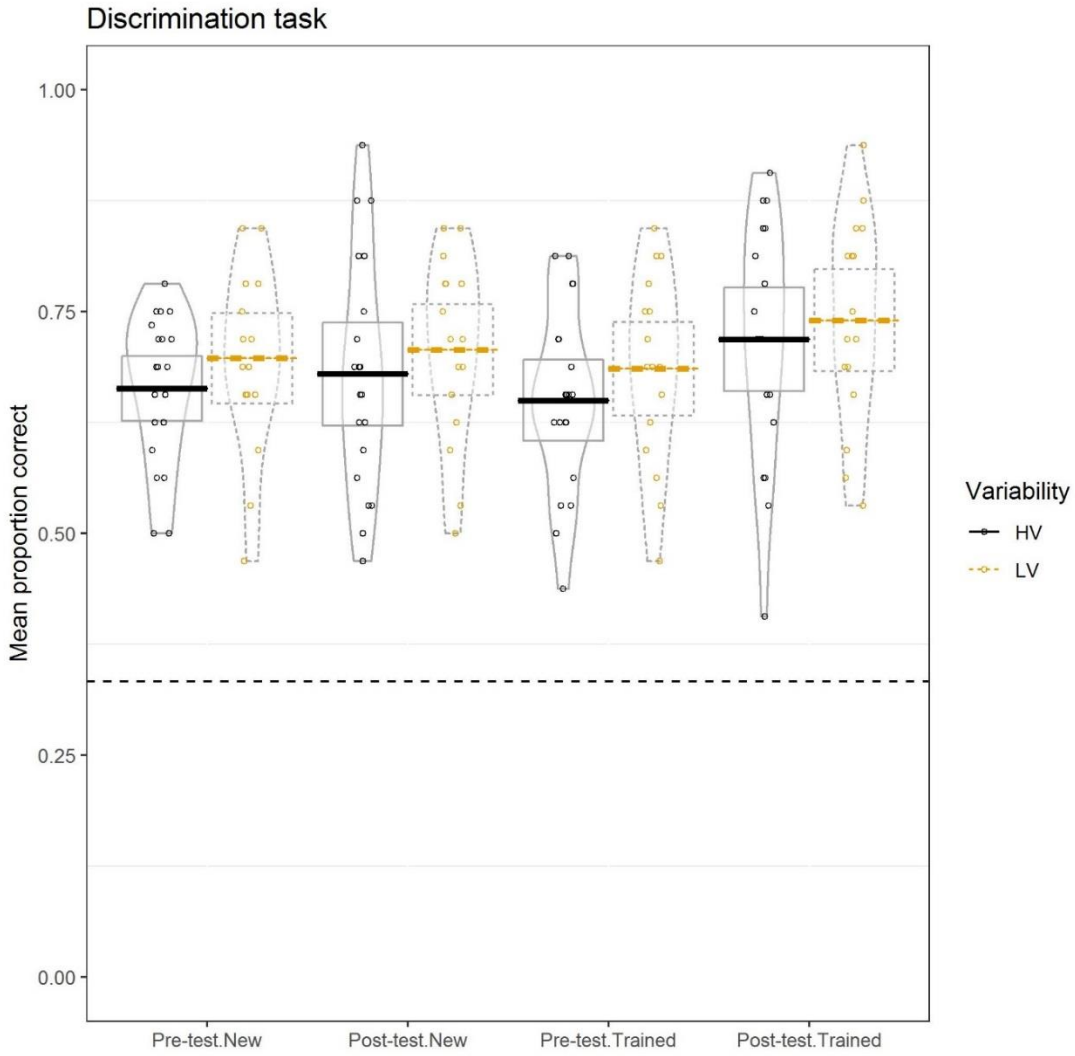


Figure 43. Accuracy results for 11-12 year olds on the pre- and post-test Discrimination task of Experiment 5, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 43 and Table 56, there was substantial evidence that 11-12 year olds performed above chance in the Discrimination task and also substantial evidence that they improved from pre- to post-test. The evidence was ambiguous for there being more improvement in HV than in LV, there being more improvement in LV than in HV, and for the hypothesis that there was greater

improvement for HV than LV for novel items only.

Final structure DiscrimG8 model: accuracy ~ session *condition*itemnovelty + VowelContrast + (session:condition:itemnovelty|participant) + (itemnovelty|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	1.689	0.133	12.680	<.001	0.833	$3.49*10^{33}$	[0, >4.8299]
improvement from pre- to post- test	session	0.219	0.070	3.143	.002	1.689	11.441	[0, >4.8299]
greater overall improvement for HV	Interaction condition: session	0.043	0.141	0.306	.760	0.219	0.672	[0.0299, 0.4299]
greater overall improvement for LV	Interaction condition: session	-0.043	0.141	-0.306	.760	0.219	0.445	[0.0299, 0.2299]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	0.020	0.303	0.067	.946	0.219	0.835	[0.0299, 0.8299]

Table 56. Mixed model results for the Discrimination task, for 11-12 year olds.

Differences between age groups

Comparing the two age groups (see Table 57), there was substantial evidence that 11-12 year-olds outperformed 7-8 year-olds overall and that they showed greater improvement from pre- to post- test. There was substantial evidence against the hypothesis of younger learners outperforming older learners, and against the hypothesis that older learners showed greater learning in HV than LV and/or younger learners showed greater learning in LV than HV.

Final structure DiscrimAgeComp model: accuracy ~
session*condition*itemnovelty*group + VowelContrast +
(session:condition:itemnovelty:group) + (session+itemnovelty+condition:group|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Older outperform younger	age	0.756	0.043	17.435	<.001	1.164	$6.15*10^{64}$	[0, >4.8299]
Older show greater improvement than younger	Interaction Session:age	0.339	0.087	3.907	<.001	2.33	417.465	[0, >4.8299]
Younger show greater improvement than older	Interaction Session:age	-0.339	0.087	-3.907	<.001	2.33	0.024	[0.0499, >4.8299]
Older improve more in HV AND/OR younger improve more in LV	Interaction Session: age: variability	-0.147	0.174	-0.844	.399	2.33	0.273	[0.2799, >4.8299]

Table 57. Mixed model results for the age comparison of the Discrimination task.

4.3.3 Orthography Identification

7-8 year olds

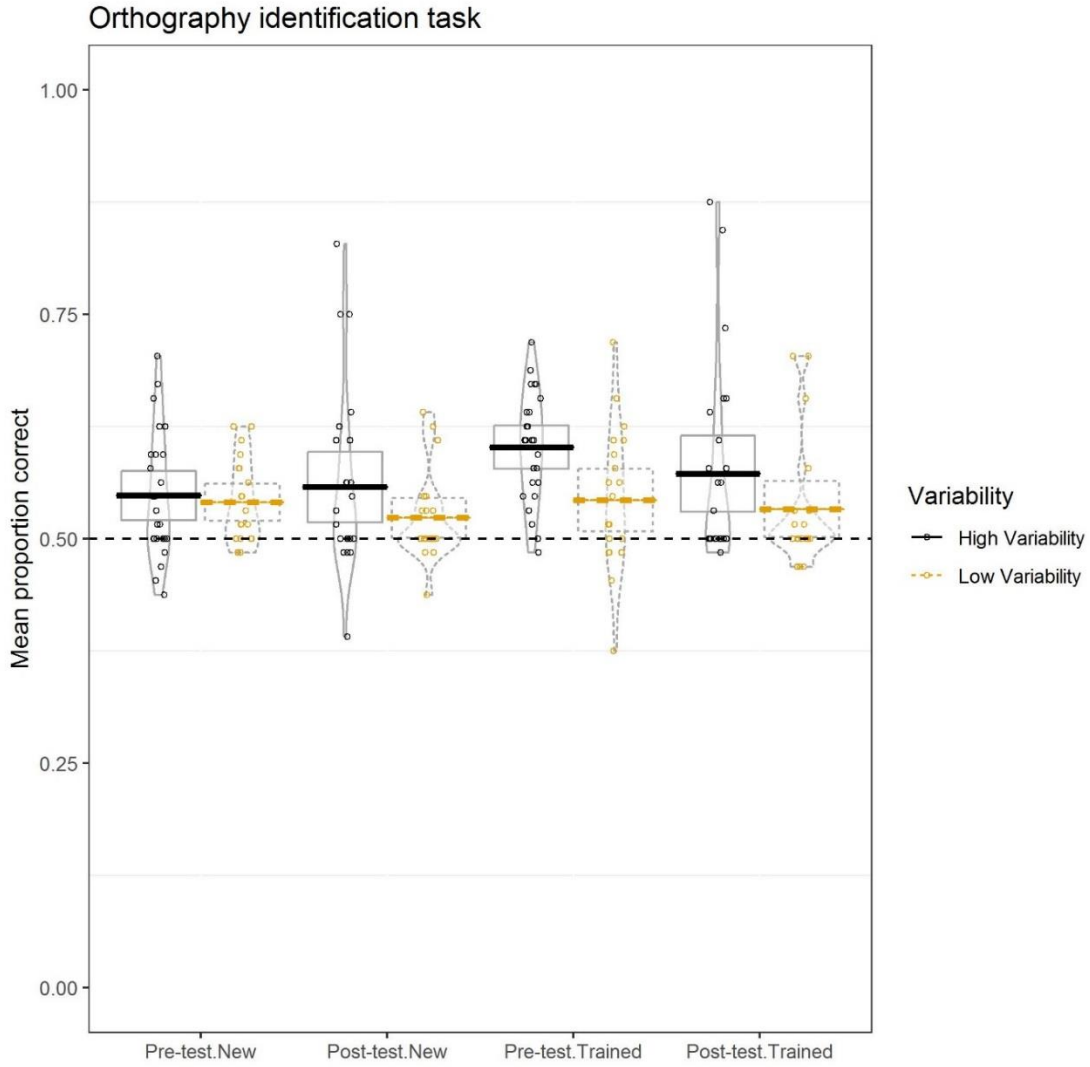


Figure 44. Accuracy results for 7-8 year olds on the pre- and post-test Orthography Identification task of Experiment 5, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 44 and Table 58, there was substantial evidence that 7-8 year olds performed above chance in the Orthography Identification task, but substantial evidence that they did not improve from pre- to post-test. The evidence was

ambiguous for there being more improvement in HV than in LV, there being more improvement in LV than in HV, and there being greater improvement in HV than LV for novel items, but no such difference for trained items.

Final structure OrthIDG4 model: accuracy ~ session*condition*itemnovelty + VowelContrast + Talker + (session:condition:itemnovelty|participant) + (condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.228	0.036	6.387	<.001	0.978	$5.12*10^7$	[0, >4.84]
improvement from pre- to post- test	session	-0.048	0.038	-1.273	.203	0.228	0.073	[0.05, >4.84]
greater overall improvement for HV	Interaction condition: session	0.013	0.076	0.176	.860	0.07	0.813	[0, 0.24]
greater overall improvement for LV	Interaction condition: session	-0.013	0.076	-0.176	.860	0.07	0.673	[0, 0.18]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	-0.186	0.155	-1.199	.230	0.023	0.865	[0, 0.19]

Table 58. Mixed model results for the Orthography Identification analysis, for 7-8 year olds.

11-12 year olds

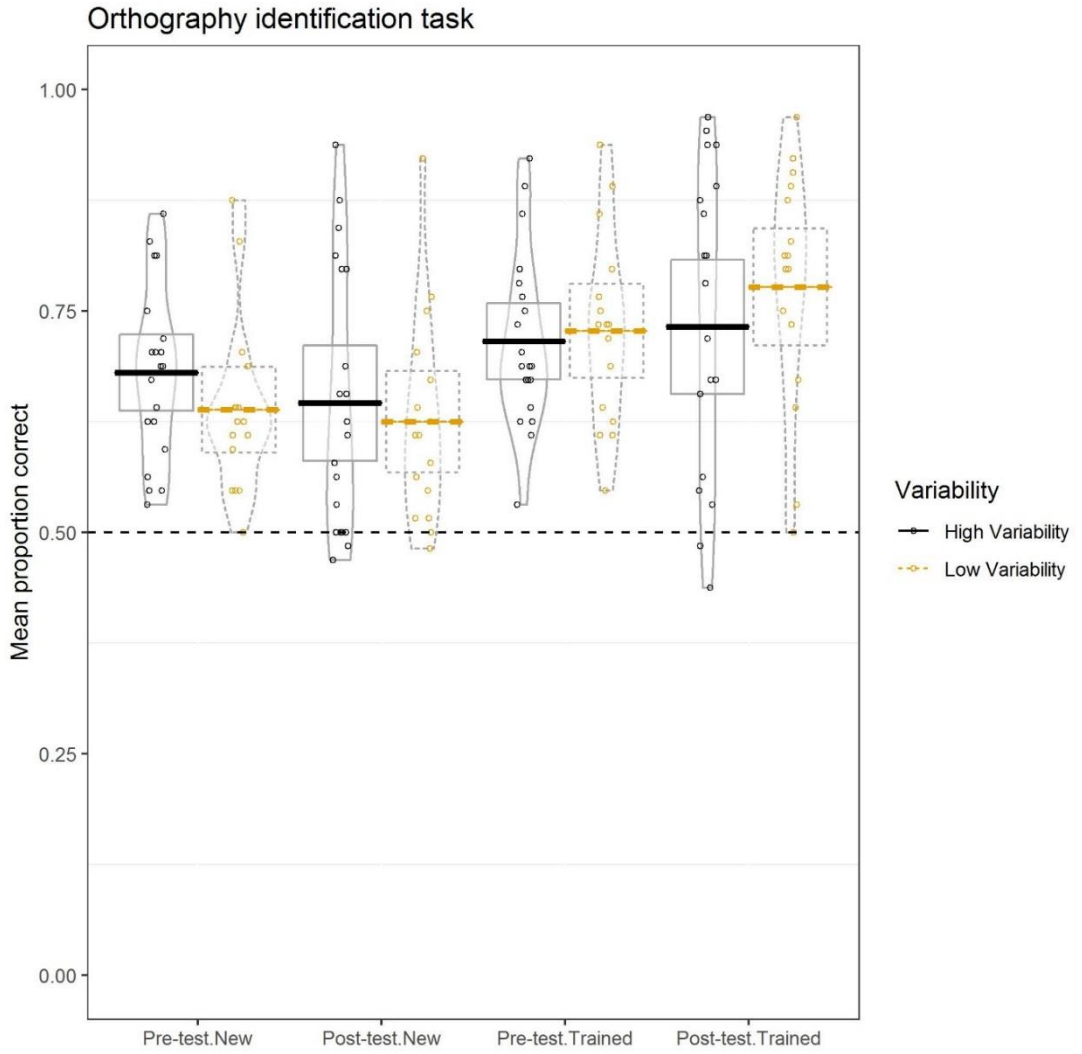


Figure 45. Accuracy results for 11-12 year olds on the pre- and post-test Orthography Identification task of Experiment 5, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

As seen in Figure 45 and Table 59, there was substantial evidence that 11-12 year-olds performed above chance in Orthography Identification, but substantial evidence that they did not improve from pre- to post-test. There was ambiguous evidence for there being more improvement in HV than in LV, and for there being more improvement in LV than in HV. There was also ambiguous evidence for there

being greater improvement for HV than LV for novel items, but no such difference for trained items.

Final structure OrthIDG8 model: accuracy ~ session*condition*itemnovelty + VowelContrast + Talker + (session:condition:itemnovelty|participant) + (session:condition:itemnovelty|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.978	0.109	9.012	<.001	0.228	$2.04*10^{14}$	[0, >4.84]
improvement from pre- to post- test	session	0.035	0.048	0.726	.468	0.978	0.099	[0.28, >4.84]
greater overall improvement for HV	Interaction condition: session	-0.168	0.100	-1.692	.091	0.035	0.639	[0, 0.1]
greater overall improvement for LV	Interaction condition: session	0.168	0.100	1.692	.091	0.035	1.57	[0, 2.33]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	-0.156	0.205	-0.764	.445	0.023	0.927	[0, 0.33]

Table 59. Mixed model results for the Orthography Identification analysis, for 11-12 year olds.

Differences between age groups

Comparing the two age groups (see Table 60), there was substantial evidence that 11-12 year-olds outperformed 7-8 year-olds. In terms of the evidence for an effect of age on pre- to post-test improvement, there was ambiguous evidence for older learners outperforming younger learners and substantial evidence against younger learners outperforming older learners. There was substantial evidence against an interaction of age and variability on pre- to post-test improvement with greater

improvement for older learners in HV condition than LV condition and/or more improvement for younger learners in LV than HV conditions.

Final structure OrthIDAgeComp model: accuracy ~ session*condition*itemnovelty + VowelContrast + Talker + (session:condition:itemnovelty|participant) + (session:condition:itemnovelty|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Older outperform younger	age	0.669	0.088	7.566	<.001	0.535	$4.10*10^{11}$	[0, >4.84]
Older show greater improvement than younger	Interaction Session:age	0.081	0.060	1.343	.179	0.356	0.729	[0, 0.81]
Younger show greater improvement than older	Interaction Session:age	-0.081	0.060	-1.343	.179	0.356	0.076	[0.08, >4.84]
Older improve more in HV AND/OR younger improve more in LV	Interaction Session: age: variability	-0.158	0.121	-1.305	.192	0.713	0.078	[0.15, >4.84]

Table 60. Mixed model results for the age comparison of the Orthography Identification task.

4.3.4 Picture Identification

Note that the analyses described here differ from the preregistration, since that document proposed analysing the data in the same way as other pre/post-tests, with item novelty as a factor. However, in hindsight, as in Experiment 2 of Study 1, the contrast between the minimal pair trials and the non-minimal pair trials is quite different from the item novelty manipulation. This is reflected in rather different performances on the minimal pair items compared to the non-pair items.

For this reason, the minimal pair items are analysed separately from the non-pair items for each age group.

7-8 year olds

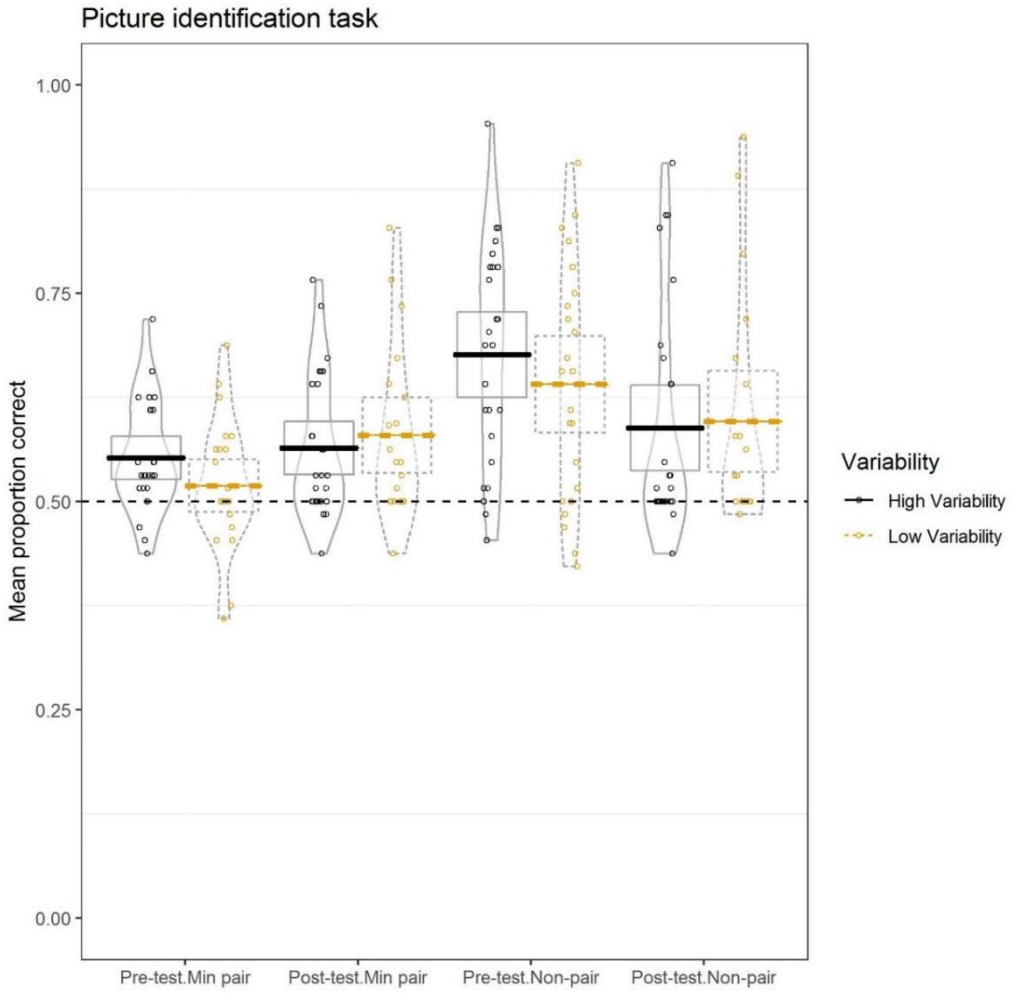


Figure 46. Accuracy results for 7-8 year olds on the pre- and post-test Picture Identification task of Study 2, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Minimal pairs

As seen in Figure 46 and Table 61, there was substantial evidence that 7-8 year olds performed above chance on the minimal pair items. The evidence was ambiguous as to there being improvement from pre- to post-test, as well as to there being more

improvement in LV training. There was substantial evidence against there being more improvement after HV than LV training.

Final structure PicIDG4minpair: accuracy ~ session*condition + VowelContrast + Talker + (session*condition||participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.224	0.050	4.507	<.001	1.373	1836.957	[0, >4.1425]
improvement from pre- to post- test	session	0.140	0.075	1.877	.061	0.767	1.075	[0.2225, 2.4625]
greater overall improvement for HV	Interaction condition: session	-0.203	0.151	-1.344	.179	0.224	0.278	[0.1825, >4.1425]
greater overall improvement for LV	Interaction condition: session	0.203	0.151	1.344	.179	0.224	1.806	[0, 2.0025]

Table 61. Mixed model results for the minimal pair trials in the Picture Identification analysis, for 7-8 year olds.

Non-pairs

As seen in Figure 46 and Table 62 there was substantial evidence that 7-8 year olds performed above chance on the non-pair items. There was substantial evidence against pre- to post-test improvement, as well as against there being more improvement after HV than LV training. There was ambiguous evidence for there being more improvement after LV than HV training.

Final structure PicIDG4nonpair: accuracy ~ session*condition + VowelContrast + Talker + (session*condition||participant) + (session:condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.572	0.079	7.253	<.001	3.277	$1.25*10^{10}$	[0, >4.1425]
improvement from pre- to post- test	session	-0.299	0.110	-2.720	.007	0.572	0.053	[0.0825, >4.1425]
greater overall improvement for HV	Interaction condition: session	-0.197	0.226	-0.871	.384	0.381	0.308	[0.3525, >4.1425]
greater overall improvement for LV	Interaction condition: session	0.197	0.226	0.871	.384	0.381	1.047	[0, 1.5625]

Table 62. Mixed model results for the non-pair trials in the Picture Identification analysis, for 7-8 year-olds.

11-12 year-olds

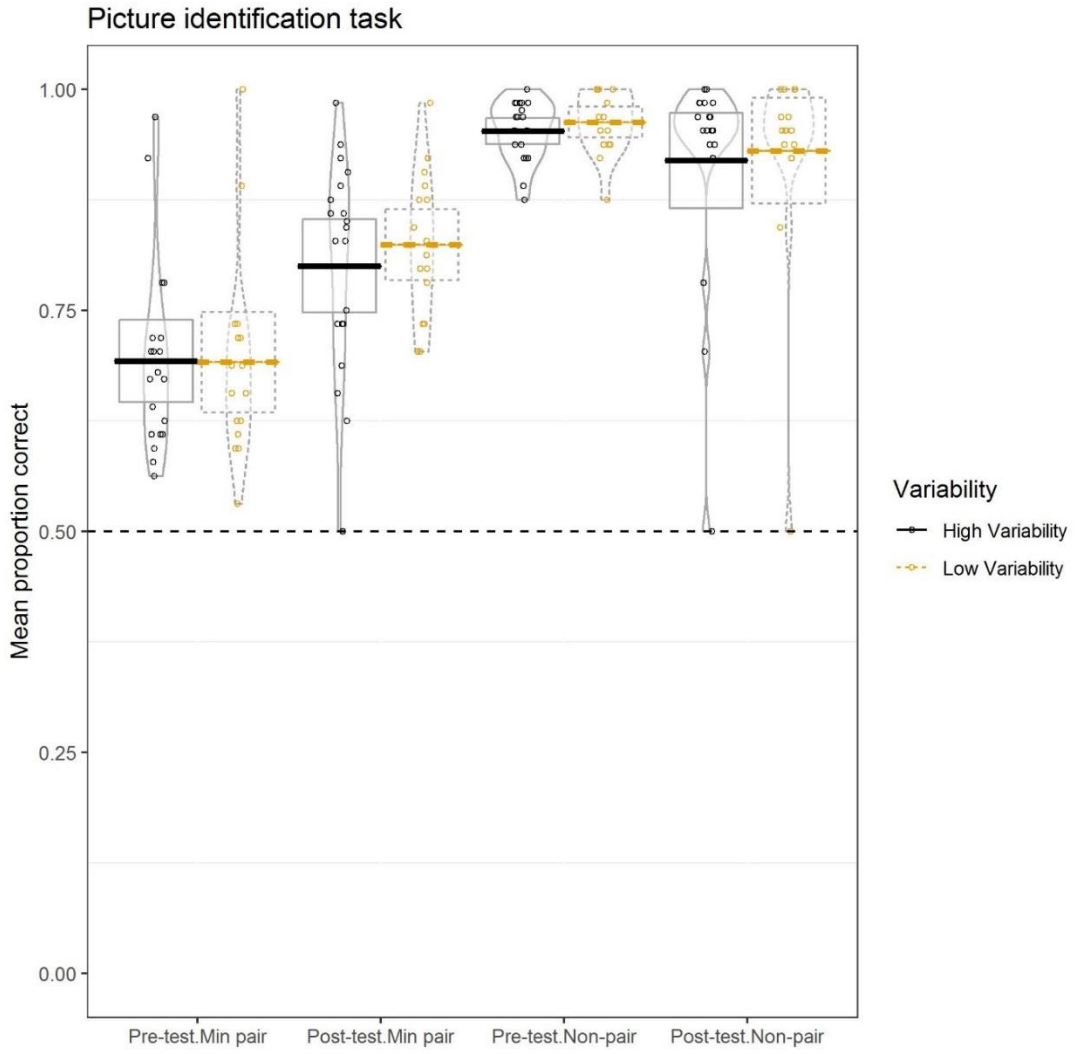


Figure 47. Accuracy results for 11-12 year-olds on the pre- and post-test Picture Identification task of Study 2, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Minimal pairs

As seen in Figure 47 and Table 63, there was substantial evidence that 11-12 year-olds performed above chance on the non-pair items, as well as for them improving from pre- to post-test. There was substantial evidence against there being more

improvement after HV than LV training, and ambiguous evidence as to there being more improvement in LV than HV training.

Final structure PicIDG8minpair: accuracy ~ session*condition + VowelContrast + (session*condition|participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	1.373	0.137	10.003	<.001	0.224	$5.46*10^{15}$	[0, >4.1425]
improvement from pre- to post- test	session	0.767	0.115	6.663	<.001	0.14	$7.05*10^{5}$	[0, >4.1425]
greater overall improvement for HV	Interaction condition: session	-0.081	0.219	-0.370	.712	0.767	0.214	[0.4625, >4.1425]
greater overall improvement for LV	Interaction condition: session	0.081	0.219	0.370	.712	0.767	0.376	[0, 0.8725]

Table 63. Mixed model results for the minimal pair trials in the Picture Identification analysis, for 11-12 year-olds.

Non-pairs

As seen in Figure 47 and Table 64, there was substantial evidence that 11-12 year olds performed above chance on the non-pair items. There was substantial evidence against pre- to post-test improvement, as well as against there being more improvement after HV than LV training and there being more improvement after LV than HV training.

Final structure PicIDG8nonpair: accuracy ~ session*condition + VowelContrast + (session*condition|participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	3.277	0.164	19.950	<.001	0.224	$2.14*10^{42}$	[0, >4.1425]
improvement from pre- to post- test	session	-0.067	0.284	-0.235	.815	3.277	0.07	[0.6625, >4.1425]
greater overall improvement for HV	Interaction condition: session	0.013	0.574	0.023	.982	2.185	0.257	[1.6525, >4.1425]
greater overall improvement for LV	Interaction condition: session	-0.013	0.574	-0.023	.982	2.185	0.248	[1.5825, >4.1425

Table 64. Mixed model results for the non-pair trials in the Picture Identification analysis, for 11-12 year-olds.

Differences between age groups

Minimal pairs

Comparing the two age groups (Table 65), there was substantial evidence that 11-12 year olds outperformed 7-8 year olds on the minimal pairs. In terms of the evidence for an effect of age on pre- to post-test improvement, there was substantial evidence for older learners outperforming younger learners, and substantial evidence against younger learners outperforming older learners. There was ambiguous evidence for an interaction of age and variability on pre- to post-test improvement with greater improvement for older learners in HV condition than LV condition and/or more improvement for younger learners in LV than HV conditions.

Final structure PicIDAgeCompminpair: accuracy ~ session*condition*group + VowelContrast + Talker + (session*condition*group|participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Older outperform younger	age	1.011	0.090	11.202	<.001	0.682	$1.58*10^{26}$	[0, >4.1425]
Older show greater improvement than younger	Interaction Session:age	0.557	0.130	4.284	<.001	0.393	2453.931	[0, >4.1425]
Younger show greater improvement than older	Interaction Session:age	-0.557	0.130	-4.284	<.001	0.393	0.056	[0.0625, >4.1425]
Older improve more in HV AND/OR younger improve more in LV	Interaction Session: age: variability	0.070	0.262	0.269	.788	0.262	0.831	[0, 0.9425]

Table 65. Mixed model results for the age comparison of the minimal pair trials of the Picture Identification task.

Non-pairs

Comparing the two age groups (Table 66), there was substantial evidence against 11-12 year olds outperforming 7-8 year olds on the non-pair items. There was ambiguous evidence for older learners outperforming younger learners and for younger learners outperforming older learners, and for an interaction of age and variability on pre- to post-test improvement with greater improvement for older learners in HV condition than LV condition and/or more improvement for younger learners in LV than HV conditions.

Final structure PicIDAgeCompnonpair: accuracy ~ session*condition*group + VowelContrast + (session*condition*group||participant) + (session:condition:group|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Older outperform younger	age	2.631	0.162	16.271	<.001	1.752	$1.8*10^{56}$	[0, >4.1425]
Older show greater improvement than younger	Interaction Session:age	-0.049	0.258	-0.191	.849	0.22	0.84	[0, 0.8625]
Younger show greater improvement than older	Interaction Session:age	0.049	0.258	0.191	.849	0.22	0.689	[0, 0.6125]
Older improve more in HV AND/OR younger improve more in LV	Interaction Session: age: variability	0.315	0.523	0.603	.546	0.44	0.574	[0, 0.9425]

Table 66. Mixed model results for the age comparison of the non-pair trials of the Picture Identification task.

4.3.5 Vocabulary

7-8 year olds

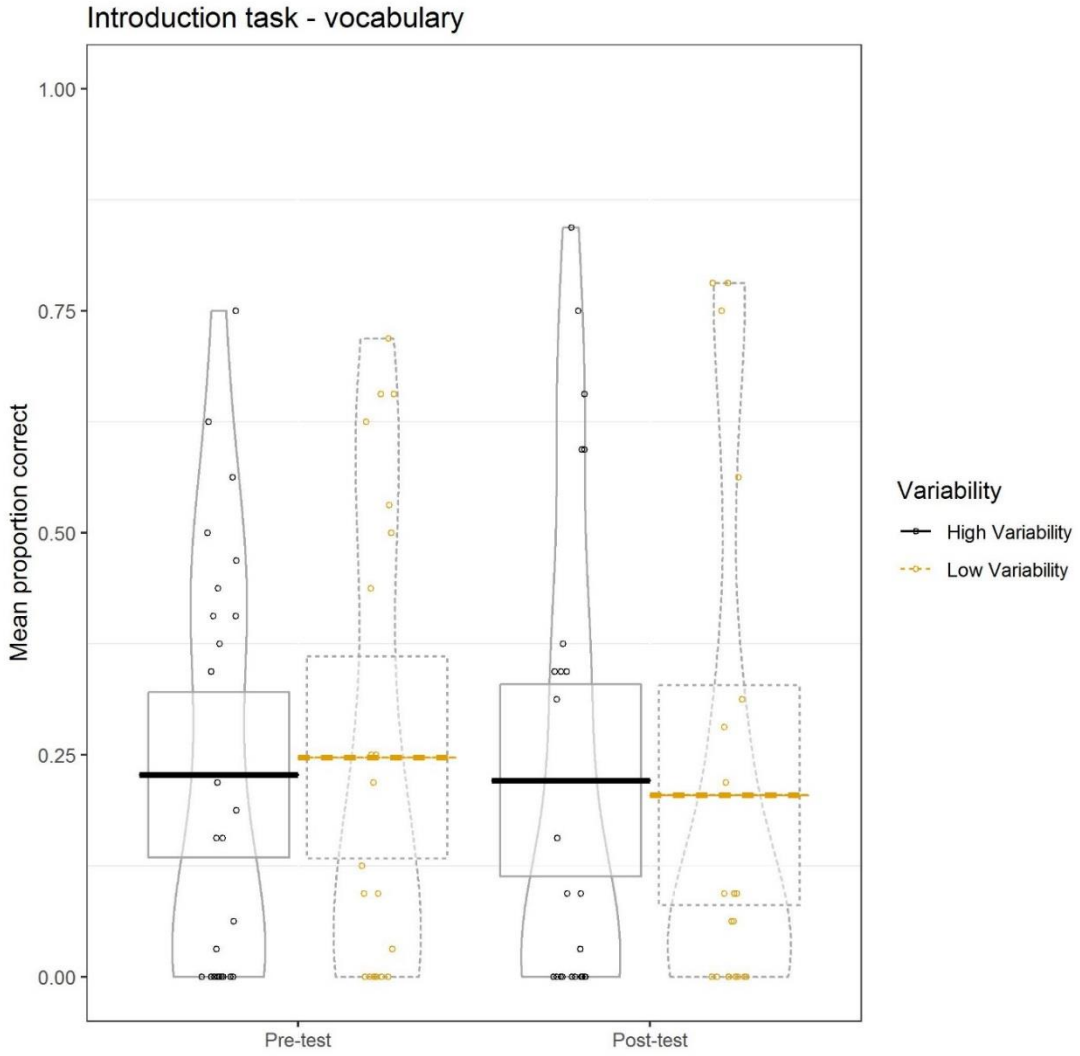


Figure 48. Accuracy results for 7-8 year olds on the Vocabulary task at pre- and post-test, comparing accuracy for HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI. There is no chance level.

As seen in Figure 48 and Table 67, there was substantial evidence against improvement from pre-test to post-test. There was ambiguous evidence for there being more improvement in HV than in LV, and for there being more improvement LV than in HV.

Final structure VocabG4 model: accuracy ~ session*condition + Vcontrast + Talker + (session*condition|participant) + (session+condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
improvement from pre- to post- test	session	-0.678	0435	-1.557	.119	1.006	0.167	[0, >6.8659]
greater overall improvement for HV	Interaction condition: session	-0.053	0.848	-0.063	.950	1.006	0.618	[0, 2.2659]
greater overall improvement for LV	Interaction condition: session	0.053	0.848	0.063	.950	1.006	0.668	[0, 2.5259]

Table 67. Mixed model results for the Vocabulary task for 7-8 year-olds.

11-12 year-olds

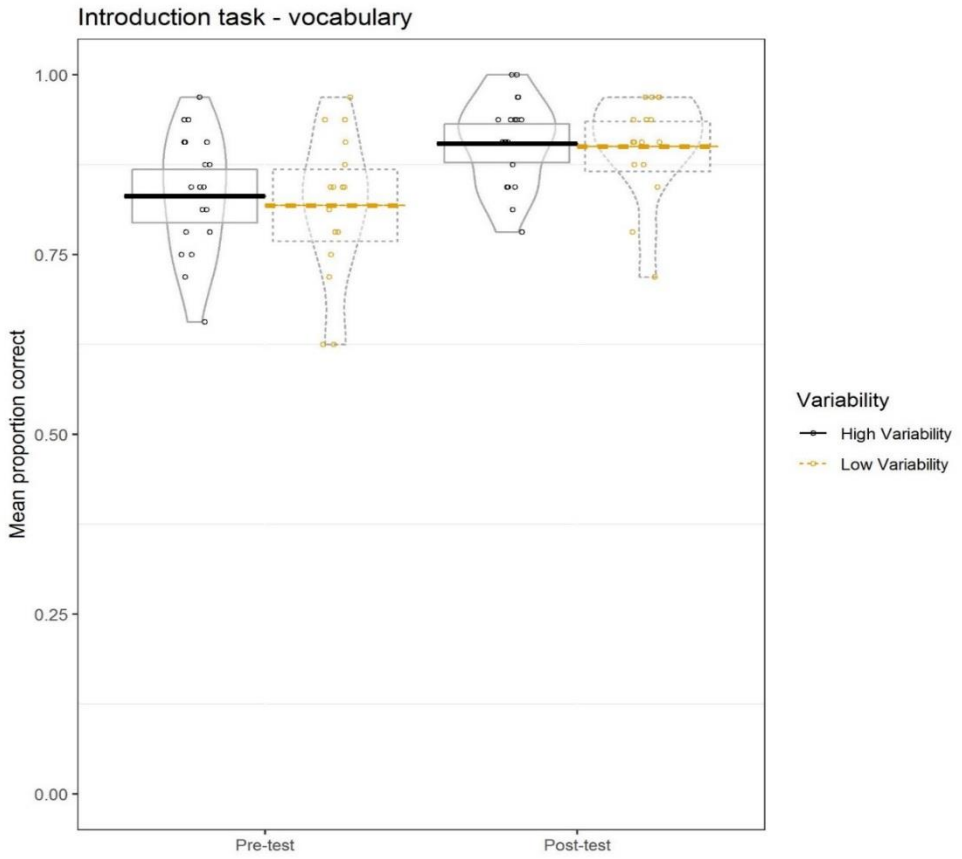


Figure 49. Accuracy results for 11-12 year-olds on the Vocabulary task at pre- and post-test, comparing accuracy for HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI. There is no chance level.

As seen in Figure 49 and Table 68, there was substantial evidence for improvement from pre-test to post-test. There was substantial evidence against there being more improvement in in HV than in LV and ambiguous evidence for there being more improvement LV than in HV.

Final structure VocabG8 model: accuracy ~ session*condition + Vcontrast + Talker + (session*condition|participant) + (condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
improvement from pre- to post- test	session	1.006	0.164	6.133	<.001	6.473	$7.36*10^6$	[0.7259, >6.8659]
greater overall improvement for HV	Interaction condition: session	-0.091	0.328	-0.276	.782	1.006	0.253	[0.7459, >6.8259]
greater overall improvement for LV	Interaction condition: session	0.091	0.328	0.276	.782	1.006	0.386	[0, 1.1759]

Table 68. Mixed model results for the Vocabulary task, for 11-12 year olds.

Differences between age groups

Comparing the two age groups (Table 69), there was substantial evidence that 11-12 year-olds outperformed 7-8 year-olds and for older learners outperforming younger learners, while there was substantial evidence against younger learners outperforming older learners. There was ambiguous evidence for an interaction of age and variability on pre- to post-test improvement with greater improvement for older learners in HV condition than LV condition and/or more improvement for younger learners in LV than HV conditions.

Final structure VocabAgeComp model: accuracy ~ session*condition*group + Vcontrast + Talker + (session*condition*group||participant) + (session*condition*group||item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
Older outperform younger	age	5.655	0.481	11.750	<.001	3.349	$6.739*10^{28}$	[0, >6.8259]
Older show greater improvement than younger	Interaction Session:age	1.266	0.387	3.272	.001	0.315	12.85	[0, >6.8259]
Younger show greater improvement than older	Interaction Session:age	-1.266	0.387	-3.272	.001	0.315	0.252	[0, >6.8259]
Older improve more in HV AND/OR younger improve more in LV	Interaction Session: age: variability	-0.296	0.766	-0.386	.700	0.21	0.893	[0, 1.6059]

Table 69. Mixed model results for the age comparison for the Vocabulary task.

4.3.6 Production – real word repetition

Rating and inter-rater reliability

Child production responses were pre-processed to cut out background noise. Inaudible responses, cases where a participant provided multiple answers, or where the audio quality was distorted were discarded altogether, resulting in the removal of 7% of the responses (211 out of 2947 recorded sound files). Sound clips were edited down to just the target production, to ensure any background noise or potential productions from other children were not mistaken for the target word. All sound files were downsampled to 22 050 samples per second, had their intensity scaled to 70 Hz, and were filtered with a band-pass filter using Praat (Boersma & Weenink, 2015). They were then used in a rating task.

The rater was a native English speaker of Standard Southern British English, who was told they would be hearing children who were learning to speak English produce some English words as well as non-words[23], and to rate each production on two levels: first, to identify which vowel the child was producing, and second to rate on a 7-point Likert scale how good of an example of that vowel the production was. The rater went through the productions in batches, and was blind to whether a trial came from the pre-test or post-test, as well as to which training condition the production came from. The rating task itself was created in Gorilla (Anwyl-Irvine et al., 2019) on the basis of Shaw et al. (2018). On each trial of the rating task, the rater heard a production from one of the child participants in the study. While this was played, there was a grid on the screen containing the 12 vowel keywords (bid, bead, bed, bird, paired, bad, bard, bud, pod, bored, hood, rude). Keywords were taken from Shaw et al. (2018). As the children's target vowels were all monophthongs, only monophthongs were presented as keywords to the rater (note that the SQUARE vowel was included in this set, as it is undergoing monophthongisation in SSBE). After replaying the production token as many times as needed, the rater clicked on the keyword with the vowel they considered to best match the target vowel in the token they had heard. Once a

[23] The rater was told there would be non-words, as mispronunciations in the children's production of the consonants were likely to lead to non-words as well as real English words being produced.

keyword was selected, the rater was asked to indicate on a 7-point Likert scale (7 = excellent, 1 = poor) how good of an example of the vowel the production they heard was. The rater rated 3401 trials in total in random order, of which only 42 trials were marked as inaudible. As control trials and to ground perception, the native SSBE speaker recordings of the stimuli files that the children were repeating were interspersed throughout the rated trials (378 control trials). Performance on these trails was generally at or near ceiling (100% accuracy for DRESS, FLEECE, LOT, STRUT, TRAP; 96% for FOOT and THOUGHT); however notably, the GOOSE vowel showed mediocre performance at 50% accuracy overall (it was rated as the FOOT vowel in 46% of the times and identified as FLEECE the other 4%). This suggests the rater had particular difficulty correctly identifying the GOOSE-vowel, and that a larger percentage of GOOSE trials will have been incorrectly rejected (false negatives), when in reality the correct vowel was produced by the children and recognised by the rater, but just mislabelled as being a different vowel. This means that the results will likely be underestimating the children's production accuracy for the GOOSE vowel. One could predict this might also have affected the rating of the other vowels, but judging from the overall near-ceiling performance on the other control trials the issue was mainly limited to correctly identifying the GOOSE vowel and not so much to incorrectly identifying other vowels to be the GOOSE vowel, meaning the influence on the result will mostly be restricted to the GOOSE vowel itself. In addition to the control trials, the rater was given 10% of the material again as a way of getting at internal consistency, for

lack of Inter-Rater Reliability with just the one rater. Intra-rater reliability was calculated using the *irr* package (Gamer, Lemon, Fellows, & Singh, 2019). Results showed that for the native controls, percentage agreement was very high at 94.7% between the first and second rating for this subset of 38 items. Cohen's Kappa showed almost perfect agreement (following criteria in Landis & Koch (1977)), κ = 0.937, which was greater than would be expected at chance, z = 13.5, p = 0. For the non-native experimental trials, percentage agreement was still high at 69.9% between the first and second rating for a subset of 249 items. Cohen's Kappa was substantial: κ = 0.662, z = 30.2, p = 0. (A table for the percentage agreement split out per vowel contrast can be found in Appendix IX, showing the GOOSE vowel did not show lower percentage agreement despite having an overall low accuracy score on the native speaker control trials). This indicates the rater had more difficulty in rating the non-native productions than the native speaker control trials.

Real word repetition

Note that due to recording difficulties described in Section 4.2.1 there are only 9 participants in the 7/8-year-olds group (6 HV, 3LV) and only 18 participants in the 11/12-year-olds group (8 HV, 10 LV). The planned and pre-registered analyses are not appropriate for the current sample size, so for this reason only a limited set of analyses will be performed which will largely remain exploratory, and for which the statistics should be interpreted with care.

Figure 50 shows the overall accuracy of the ratings - i.e. the proportion of trials for which the rater accurately identified the vowel the children produced, in each condition pre and post-test. There is evidence for an improvement from pre-test to post-test and this was confirmed statistically (main effect of session for 7-8 year-olds: β = 0.551, SE = 0.185, z = 2.976, p = .003; $BF_{H(0,0.36)}$[24] = 30.063 [RR 0.7885, >6.9857]. Main effect of session for 11-12 year-olds: β= 0.356, SE = 0.102, z = 3.495, p <.001; $BF_{H(0,0.55)}$[25] = 133.208 [RR 0.7885, >6.9885]).

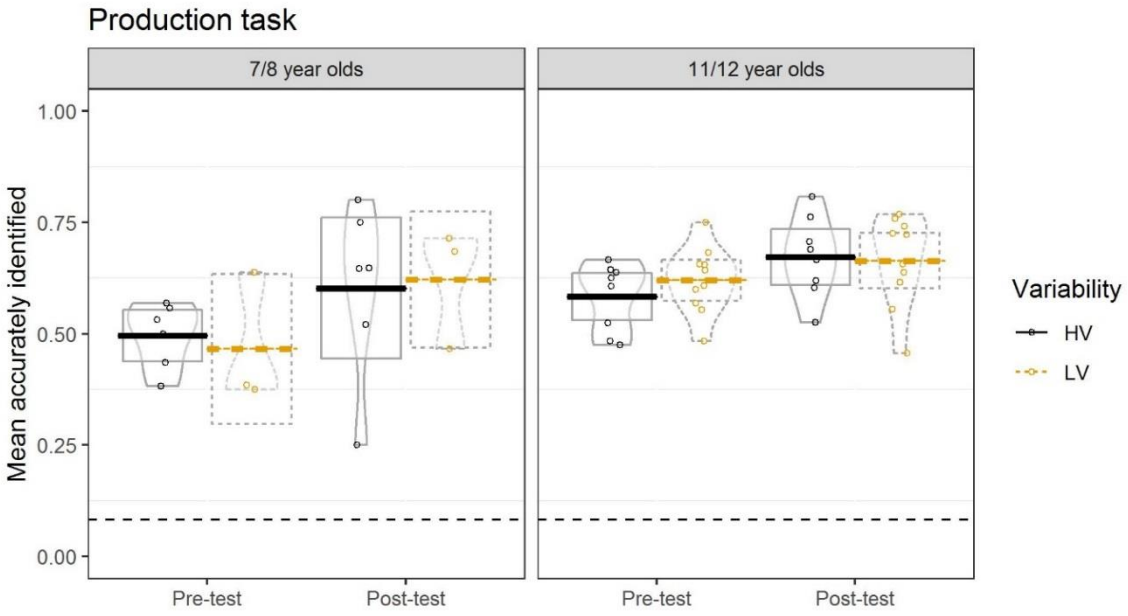


Figure 50. Binary rating accuracy results for 7-8 year-olds and 11-12 year-olds on the Production task at pre- and post-test, comparing accuracy for HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI. Chance is at 1/12.

[24] BF based on the equivalent effect in the ProdG8 model.
[25] BF based on the equivalent effect in the ProdG4 model.

However, this rating accuracy score looks only at whether the rater identified the right vowel (i.e. the *hits*), without taking the trials incorrectly indicated as being that target vowel (i.e. the *false alarms*) into account. In this case it is exactly those incorrect trials that reflect what is going on in children's vowel production. A measure which can capture this is the accuracy prime score: it takes into account not just how likely the rater was to correctly identify a child's production (e.g. how likely the rater was to choose the target keyword upon hearing the child's production), but it also penalises the accuracy score when target vowels are identified as incorrect keywords instead. The more positive the accuracy prime score is, the more likely the rater was to correctly identify the target vowel and, therefore, the better the children's production of that vowel. If the accuracy prime score is negative, the rater was more likely to identify the children's production of that vowel as an incorrect keyword, and thus the child's production maps onto a vowel that is not its SSBE target (e.g. their attempt at TRAP might sound more like SSBE DRESS). In this case, the accuracy prime score is also weighted: without this weighting, the score would assume all response options were equally likely. However, the vowels produced here are not all equally confusable: some are more likely to be confused than other vowels (e.g. TRAP is more likely to sound like DRESS than it is to sound like FLEECE) and thus are more likely candidates throughout the task. To account for this, the average proportion at which the rater chose a particular vowel throughout the entire experiment is taken into account on the individual trials.

The prime accuracy data are shown in Figure 51 and Figure 52. Note that, unlike for previous analyses, plots are separated by vowel given that large by-vowel differences are expected here (however since no effect of item status was found, these figures are collapsed over trained and untrained items for readability). For the analyses however, the same hypotheses are investigated as before and these are shown in Table 70 and Table 71. It can be seen that, contrary to the binary accuracy rating, there was no improvement from pre-test to post-test for either age group with evidence for the null for 7-8 year-olds (though only just), and ambiguous evidence for 11-12 year-olds. There are also no interactions between session and input variability or item novelty; for 7-8 year-olds there is ambiguous evidence for both the interaction between condition and session, and the interaction between condition, session, and item status, while for 11-12 year-olds the evidence for an interaction between session and condition is ambiguous, while there is evidence for the null for the interaction between session, condition, and item status. This suggests that while the children's production might have become more accurate purely in terms of hits as visualised in the binary data, the number of false alarms has also increased, evening out the overall improvement.

Final ProdG4 model structure: lmer(accuracy_prime ~ session*condition*status + Vowel + Talker + (session*condition*status||participant) + (1|target)

Hypothesis	fixed effect in model	beta	SE	df	t	p	predicted effect x	Bayes factor	Robustness Region
improvement from pre- to post- test	session	0.175	1.320	6.561	0.132	.899	4.195	0.331	[4.1833, 4.1933]
greater overall improvement for HV	Interaction condition: session	-2.238	2.774	6.557	-0.807	.448	4.195	0.345	[0, 4.3633]
greater overall improvement for LV	Interaction condition: session	2.238	2.774	6.557	0.807	.448	4.195	1.035	[0, 17.7133]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	-1.172	3.312	3.176	-0.354	.746	4.195	0.501	[0, 7.0433]

Table 70. Linear mixed model results for the Production task, for 7-8 year-olds. Note: there are only 9 participants in the 7-8 year-olds group (6 HV, 3LV), so the statistics should be interpreted with care.

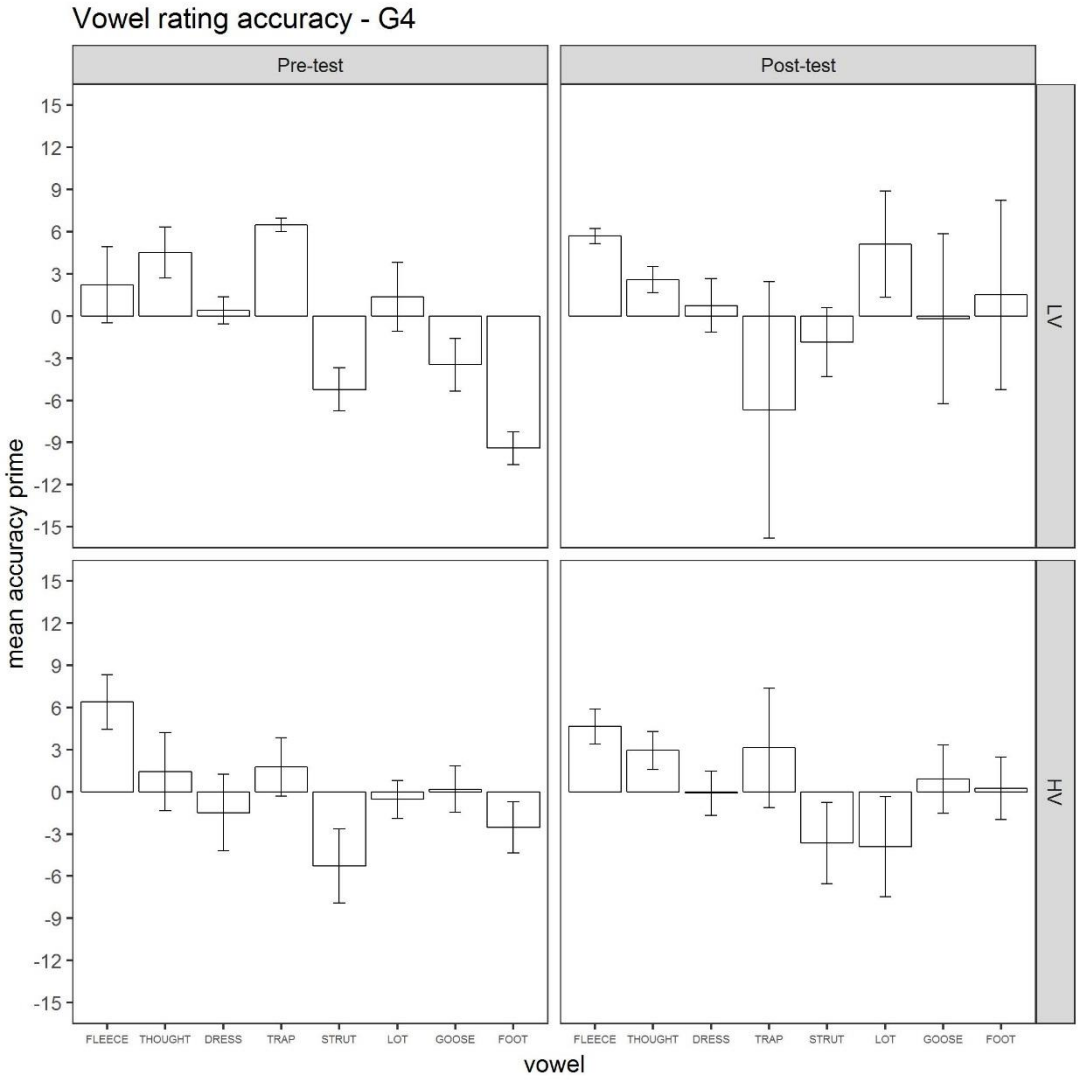


Figure 51. Mean Accuracy' scores (error bars = 95% Confidence Intervals) obtained for categorization ratings of each vowel in the 7-8 year-olds in the Production task at pre- and post-test, comparing accuracy for HV versus LV training input. A negative score means the rater more often selected an incorrect keyword than the correct target for that vowel.

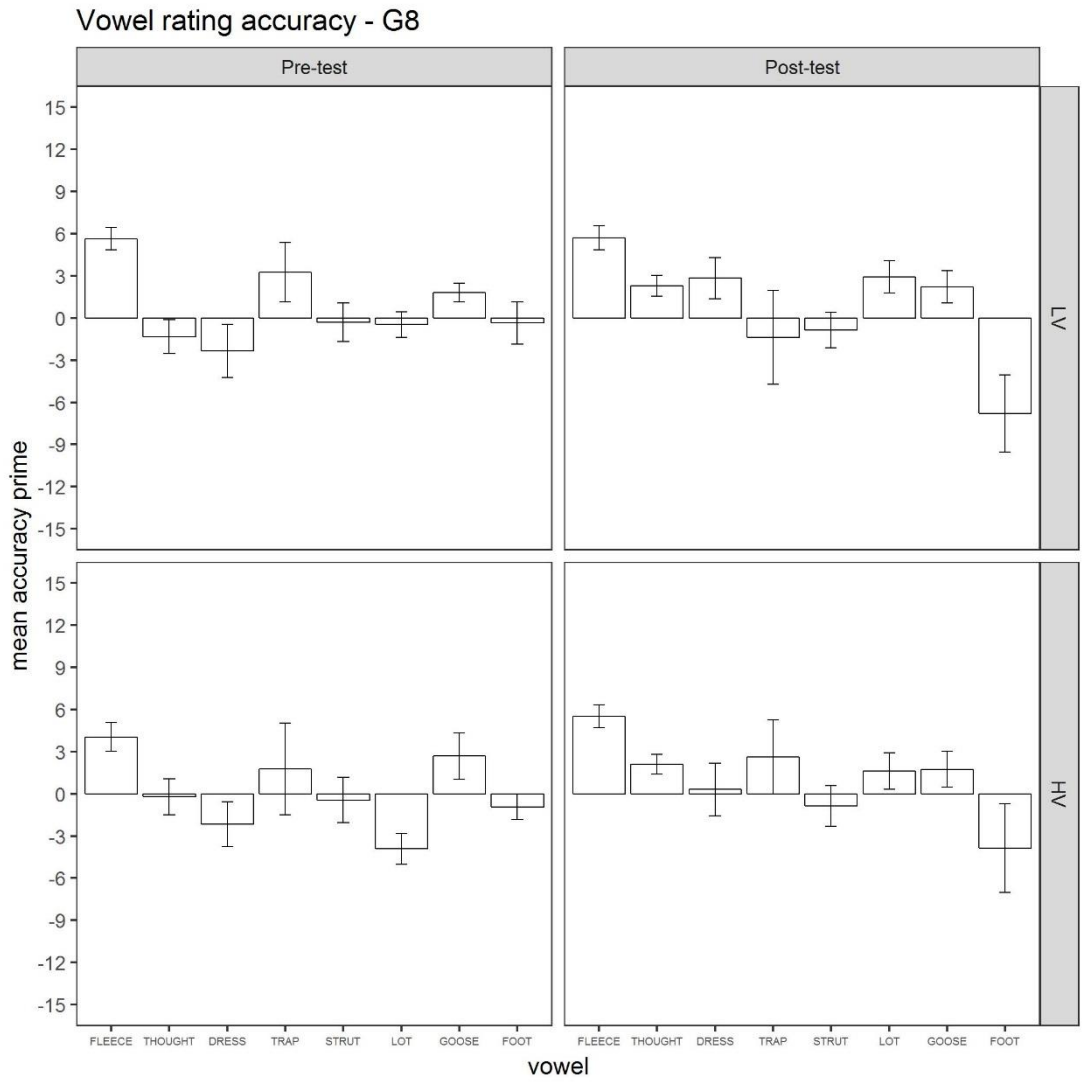


Figure 52. Mean Accuracy' scores (error bars = 95% Confidence Intervals) obtained for categorization ratings of each vowel in the 11-12 year-olds in the Production task at pre- and post-test, comparing accuracy for HV versus LV training input. A negative score means the rater more often selected an incorrect keyword than the correct target for that vowel.

Final ProdG8 model structure: lmer(accuracy_prime ~ session*condition*status + Vowel + Talker + (session*condition*status||participant) + (status+condition|target)

Hypothesis	fixed effect in model	beta	SE	df	t	p	predicted effect x	Bayes factor	Robustness Region
improvement from pre- to post- test	session	0.537	1.036	15.949	0.519	.611	3.84	0.41	[0, > 4.7533]
greater overall improvement for HV	Interaction condition: session	0.914	2.084	15.932	0.439	.667	3.84	0.67	[0, 8.8233]
greater overall improvement for LV	Interaction condition: session	-0.914	2.084	15.932	-0.439	.667	3.84	0.362	[0, 4.1733]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	-0.195	1.158	2016.794	-0.168	.866	3.84	0.257	[0, 2.8733]

Table 71. Linear mixed model results for the Production task, for 11-12 year olds. Note: there are only 18 participants in the 11/12-year-olds group (8 HV, 10 LV), so the statistics should be interpreted with care.

When looking at the figures above, it can be seen that some vowels get more accurate while others get substantially less accurate from pre-test to post-test. The poorly identified GOOSE vowel will have resulted in lower accuracy prime scores for that vowel in particular, but will not have had a strong effect on the overall results: the robustness regions for the Bayes Factors in each of the tables show that none of the results are very close to changing the overall direction or interpretation of the effects. To illustrate what is going on in children's production in a more qualitative way, the confusion matrix in Table 72 shows which vowel keywords the children's productions of the target vowels were most often identified as. Most confusions are very much in line with the phonetic descriptions from Chapter 2.

TRAP most often gets confused with DRESS (though note the reverse hardly happens), and it also gets confused with STRUT; for 11-12 year-olds TRAP is more consistently confusable with DRESS than it is for 7-8 year-olds. LOT gets confused with STRUT, while STRUT gets confused with TRAP (particularly for 7-8 year-olds). This confusion pattern suggests there is a difference in the production of TRAP by 7-8 and 11-12 year-olds.

7-8 year-olds								
Rater response	Target vowel							
	DRESS	FLEECE	FOOT	GOOSE	LOT	STRUT	THOUGHT	TRAP
DRESS	70.59	0.00	6.94	1.25	0.00	4.60	1.45	20.73
FLEECE	0.00	88.57	1.39	21.25	1.20	3.45	1.45	1.22
FOOT	2.35	0.00	48.61	11.25	4.82	3.45	18.84	0.00
GOOSE	1.18	5.71	5.56	38.75	0.00	1.15	2.90	0.00
LOT	5.88	1.43	9.72	8.75	60.24	5.75	24.64	2.44
STRUT	1.18	0.00	2.78	0.00	24.10	52.87	4.35	14.63
THOUGHT	0.00	1.43	6.94	8.75	1.20	1.15	46.38	0.00
TRAP	2.35	0.00	0.00	0.00	2.41	18.39	0.00	48.78
KIT	4.71	2.86	16.67	10.00	2.41	0.00	0.00	1.22
NURSE	4.71	0.00	0.00	0.00	0.00	2.30	0.00	2.44
SQUARE	7.06	0.00	1.39	0.00	0.00	0.00	0.00	3.66
START	0.00	0.00	0.00	0.00	3.61	6.90	0.00	4.88
11-12 year-olds								
Rater response	Target vowel							
	DRESS	FLEECE	FOOT	GOOSE	LOT	STRUT	THOUGHT	TRAP
DRESS	80.84	0.00	5.08	0.39	0.00	1.89	0.39	30.80
FLEECE	0.77	87.25	1.95	7.06	0.00	0.00	0.39	0.00
FOOT	2.68	0.80	55.47	22.75	2.34	4.17	15.56	1.14
GOOSE	0.00	3.59	2.73	45.88	0.39	0.38	1.56	0.00
LOT	1.53	0.40	11.72	4.71	60.55	4.55	19.46	1.52
STRUT	2.30	0.00	5.86	0.00	28.91	75.00	1.95	6.08
THOUGHT	0.38	1.99	9.38	16.47	1.56	0.00	57.59	0.00
TRAP	3.83	0.00	0.39	0.00	1.56	9.47	1.17	46.77
KIT	4.98	5.98	4.30	2.75	0.00	0.38	0.00	0.00
NURSE	0.77	0.00	3.12	0.00	0.39	1.52	1.17	3.80
SQUARE	1.92	0.00	0.00	0.00	0.00	0.38	0.00	6.84
START	0.00	0.00	0.00	0.00	4.30	2.27	0.78	3.04

Table 72. Confusion matrix showing the percentage of rater vowel keyword responses for the target vowels produced by 7-8 year-olds and 11-12 year-olds.

Overall, FOOT and GOOSE seem less consistent in what vowels they are confused with, though for 11-12 year-olds GOOSE is more often confused with FOOT than for 7-8 year-olds (though again, note the rater was generally less consistent in the

rating of the GOOSE vowel). Surprisingly, THOUGHT was not as stable in its production as would be expected from the perceptual data: it gets confused with FOOT and LOT for both age groups, though is slightly more stable for 11-12 year-olds. This confusion of the long vowel THOUGHT with short vowels FOOT and LOT suggests the children might have had particular trouble in producing the correct vowel length.

For the vowels that were accurately identified, the goodness rating was also investigated (see Figure 53). Most vowels are rated around the middle of the rating scale, but productions of FOOT, GOOSE, and THOUGHT seem to be rated slightly lower compared to the other vowels. When looking at the statistics, both age groups showed substantial evidence against improvement from pre- to post-test (7-8 year olds: $\beta = 0.170$, SE = 0.178, $t(7.432.) = 0.953$, $p = .371$; $BF_{H0(0, 2.68)}$[26] = 0.175, RR[1.3833, >10.5833]; 11-12 year-olds: $\beta = -0.124$, SE = 0.101, $t(16.898) = -1.222$, $p = .238$; $BF_{H0(0,2.24)}$[26] = 0.023, RR[0.7033, >10.5833]), indicating there was no change in the goodness ratings as a result of training.

[26] BF based on the maximal possible difference: maximum score on the scale - score at pre-test per group.

Final GoodnessG4 model: lmer(Goodness ~ condition.ct*session.ct*status.ct + target vowels + (condition.ct*session.ct*status.ct||participant) + (session.ct|target)

Final GoodnessG8 model: lmer(Goodness ~ condition.ct*session.ct*status.ct + target vowels + (F1_VS_F2 + F1_VS_F3) + (condition.ct*session.ct*status.ct||participant) + (session.ct|target)

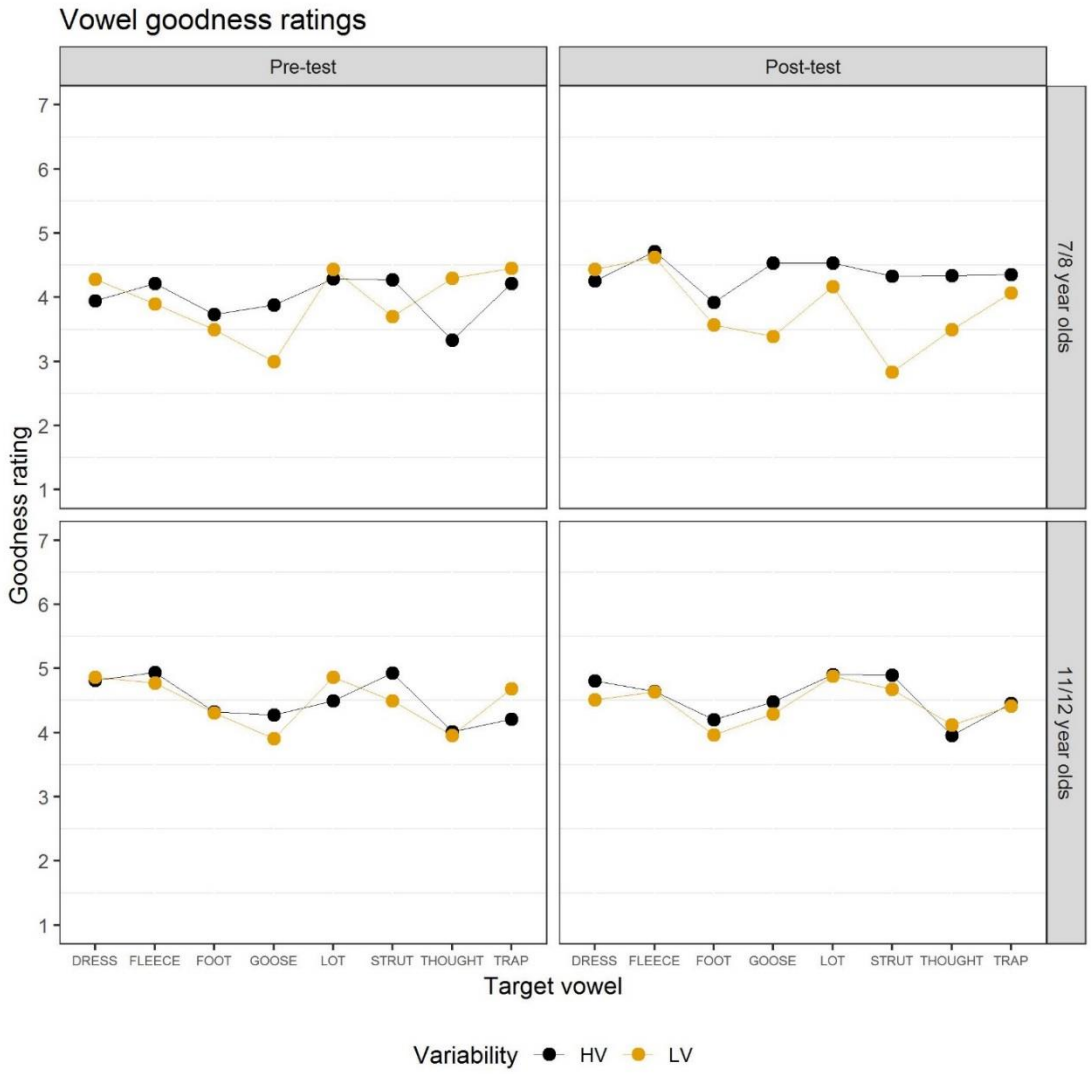


Figure 53. Rater vowel goodness rating results for 7-8 year-olds' and 11-12 year-olds' spoken word production for the pre- and post-test Production task, comparing vowel goodness for HV versus LV training input.

4.4 Discussion

The present study investigated the role of input variability in children acquiring L2 vowels through means of a phonetic training paradigm. Key questions were whether there was improvement as a result of training, whether input variability affected performance, and whether these two factors were different for younger and older children.

Overall, this study showed evidence for improvement across training in both groups, although 11-12 year-olds improved more than 7-8 year-olds. However, 7-8 year-olds did not show substantial evidence of improvement in *any* of the tests, with evidence for the null in every case except for the minimal pair Picture Identification task (i.e. the task most like training), where the evidence for improvement was ambiguous, and the Production task, where the sample was small. In contrast, 11-12 year-olds improved in Discrimination, minimal pair Picture Identification, and Vocabulary. They did not improve in the non-pair Picture Identification or Orthography Identification, with evidence for the null in each case, though note they were at ceiling in the former task. Evidence of improvement was ambiguous for the Production task. Importantly, despite showing improvement in three of the tests, 11-12 year-olds did *not* show evidence for a variability benefit in generalisation in any task, with evidence for the null in the case of Picture Identification and Vocabulary. The only task that showed any evidence in terms of a variability effect is the Training task, where the 11-12 year-

olds did better with LV training than HV training as predicted. There was no evidence that variability affected the 7-8 year-olds and 11-12 year-olds differently.

The following section will first consider why improvement was found in some tasks, but not others, before going into more detail on not finding a variability effect.

4.4.1 Effects of training on test performance

Both age groups showed improvement during training. However, improvement from pre- to post-test is a different matter, with only 11-12 year-olds improving in any task. It is very surprising that, although the 7-8 year-olds do show improvement during training itself, indicating learning in this task, they do not show any improvement on the pre/post-test tasks. Note that they are above chance *overall* on all of the tasks which have a measure of chance (Discrimination, Picture Identification and Orthography Identification), 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, they do not seem to be able to generalise their trained abilities to novel voices (as were used in every test), although note evidence for generalisation in the minimal pair Picture Identification, the task most like Training, is ambiguous. This lack of generalisation goes against previous literature, where child learners of the same age were able to extend their knowledge from

phonetic training to novel speakers (Giannakopoulou et al., 2017, 2013; Shinohara & Iverson, 2013). Potential explanations for this lack of improvement could lie in the children's ability to deal with the task demands: those studies where younger children could generalise to novel speakers and items tended to only focus on one phoneme contrast, while this study taught them four vowel contrasts at the same time. It is possible that the younger learners might require longer training to show generalisation: even after 8 training sessions, their performance did not reach the starting point of the 11-12 year-olds in this study. Evidence that initially weaker learners may eventually "catch up" with additional training comes from Iverson & Evans (2009) who found that while adult German learners improved more quickly on English vowel contrasts than adult Spanish learners, the Spanish learners caught up to the same level of improvement and generalisation after additional training sessions.

Another thing that might have influenced the 7-8 year-olds' performance is their seemingly poor vocabulary knowledge. Part of this poor performance is due to the task demands: the vocabulary results reported above are answers to an open question requiring children to provide the Dutch word meaning; the younger children might not have had the skills required to complete the task. The familiarity task briefly mentioned but not further reported earlier (see footnote 22 on page 260) does show that the 7-8 year-olds say they are familiar with about 60% of the word meanings at pre-test, and this goes up to about 75% at post-test. However, given that at post-test they have been exposed to those same words for

two weeks, this score is still rather low, indicating this familiarity task is not entirely getting at their vocabulary knowledge either. However, regardless of whether the vocabulary task was the best way to measure the children's vocabulary knowledge, not having full knowledge of the vocabulary used in training might have influenced how well they performed in the tasks. Since the task required participants to link spoken words to images with orthography underneath it in training, children who had a better vocabulary knowledge will have found this easier and rather than focussing on figuring out the word meaning, they will have been able to fully focus on the phonetic contrast they were being trained on.

Three tasks in the current study did not show any improvement from pre- to post-test, even in 11-12 year-olds: non-minimal pair Picture Identification, Orthography Identification, and Production, with evidence for the null in the former and ambiguous evidence in the latter two. For the non-minimal pair Picture Identification task, this null result is clearly a ceiling effect (especially since this is a measure of vocabulary learning, and improvement is found for this age group in the other Vocabulary test). However, the fact that there is evidence of no change in the Orthography Identification task may indicate that participants are not using the orthography presented during training when learning the vowel contrasts. This is surprising given Escudero et al. (2008), who found that for Dutch adult learners of English, spelling was especially helpful initially when learning words relying on non-native phoneme contrasts. Since the training task used in this study also involves learning vocabulary that relies on being able to distinguish non-native

minimal pairs, orthography was expected to be helpful for the children in this study as well. However, despite this, it seems that they may not have made any use of the spelling in the training task. This could in part be due to the opacity of the English spelling system; even though the stimuli chosen here were created to be as transparent as possible, the grapheme-to-phoneme mapping was less consistent in the GOOSE-FOOT and FLEECE-THOUGHT contrasts, while DRESS-TRAP and STRUT-LOT where entirely consistent. This may have resulted in more difficulty in the Orthography Identification task for those particular contrasts, which means using the spelling in the Training task might have been detrimental or might have taken up too much processing power. It is possible that children may have already been exposed to enough English in their daily life to have learnt that English orthography is not as reliable a cue as to how the word ought to be pronounced, or at least less of a reliable cue than spelling in Dutch. At the same time, their familiarity with Dutch orthography might mean that this knowledge might interfere with their ability to make use of the English orthography: English orthography might lead to confusion particularly for those spellings that might activate a Dutch vowel that is distinctly dissimilar to the English target. Another possibility is that children might not be accustomed to using orthography as a source of information at all yet, even in their L1. The previous studies investigating the role of orthography have tested adults, who have many years of print exposure and have automatised the process of extracting information from orthographic cues. Children, especially the younger group, are developing readers in their L1

still, and are thus much less used to using orthography in general. They might not have automatised or even fully acquired the skill to decode information from the orthography; processing the orthographic information might then be too costly still compared to the value of the information they could get from it, or it might even be distracting. They might instead benefit more from relying on other cues, such as pictures in this case, which are less costly to process. The idea that younger children have more difficulty with orthography is in line with the findings of Shinohara (2014), where these children did not improve at all.

Similarly to the Orthography Identification task, the Production task did not show any evidence of improvement from pre- to post-test, at least when using the prime score measure which takes into account the number of false alarms as well as the correct answers. This is despite there being a sizeable literature demonstrating perception training can transfer to production. The original seminal studies investigating the feasibility of HVPT suggested perceptual learning could transfer to production in adults after training Japanese learners on English /l/-/r/ (Bradlow et al., 1997, 1999). Similar results were also found more recently for Japanese children who were trained on the English /l/-/r/ contrast (Shinohara & Iverson, 2015), and for Spanish children learning English /i/-/ɪ/ (Evans & Martín-Alvarez, 2016), suggesting that perception training alone might be enough to at least partially improve phoneme production. The findings of the current study are, therefore, slightly unexpected. Note, however, that the evidence is ambiguous and that the production analyses in the current study were based on a subset of the

total number of participants, as technical difficulties resulted in data loss. It is quite likely that a larger sample is needed to find group level improvement in the Production task. Additionally, it is possible that the changes in production are fairly subtle and cannot be easily captured in the rater identification task. For example, it is possible that participants might be producing *covert contrasts* : the children might be producing the different vowels in a contrastive way that does not align with the SSBE contrast, resulting in the contrast being perceived as homophonous to native SSBE speakers (see Scobbie, Gibbon, Hardcastle, & Fletcher (2000) for a case study). For example, children might be producing the DRESS-TRAP distinction by producing a length distinction, e.g. /bed/-/be:d/, rather than a distinction in vowel quality. Since a length distinction is not a relevant cue to the SSBE /e/-/æ/ contrast, a native speaker might not perceive the two productions as different vowels, but as variants of the same vowel instead. In isolation, these productions could be misidentified and would not be very good examples of their target vowel, but in conversation they could be perfectly functionally contrastive. The current study does not allow for testing this, so further investigation using different methods would be required.

Given the complete lack of improvement found in 7-8 year-olds, one thing to consider is whether the test tasks were entirely age-appropriate for this group. Since the younger children did perform above chance and show improvement in training (as was the case in Chapter 3), the task used in training seems suitable

enough for both the 7-8 year-olds and the 11-12 year-olds. Nevertheless, the full length of training might have meant that towards the end of the study, they may have experienced decreased motivation. Performance towards the end of each session might also have been lower given that younger children have a shorter attention span. However, this could only partially explain their poor performance on the post-test tasks: since the post-test took place on a separate day from training, a shorter attention span should not have influenced performance for the initial post-test tasks, and yet performance for 7-8 year-olds is consistently low across the entire task battery. This leads to the possibility that the task demands for the post-test tasks were too high for the younger children: in particular, the response mode might have been too difficult for them. While the Picture Identification task and Discrimination task should have been appropriate given the 7-8 year-olds performance on identical tasks in Study 1, not all tasks might have been as suitable for the age group. The Orthography Identification task might have asked too much of the children, relying on orthographic knowledge they were still developing and thus might not have been able to access yet (this will be discussed further in Section 5.1 below). Similarly, the Vocabulary Task required open ended responses that the children had to type in; as noted in the Procedure above in Section 4.2.4, 7-8 year-olds struggled with this task and this will have impeded their performance. This suggests that some of the task demands might at least have played a role in the poorer performance and lack of improvement found in the 7-8 year-olds, even if it cannot explain the consistent lack of improvement across the entire test battery.

Finally, it is worth noting that for the first time in this thesis, pre- to post-improvement was found in the Discrimination task, although for the 11-12 year-olds only. Recall that in Study 1, no improvement was found in either the child or adult participant groups for any of the experiments, which was attributed to the one-session training length. The longer exposure in the current study seems to have been helpful for the older children to be able to improve their discrimination abilities. The lack of discrimination improvement in younger children is consistent with their performance on the other pre/post-tasks, suggesting it reflects more general difficulty in generalising from Training.

4.4.2 Variability

Returning to the central question of this thesis, as in the previous chapter, this study found no evidence of an HV benefit. Recall from Section 1.3.5, that there are theoretical reasons to expect a variability benefit for generalisation, since encountering variability should help listeners to learn to ignore speaker-specific cues and tune in to the relevant cues for the phonetic contrasts. In Chapter 3, the lack of variability effect was ascribed to potentially resulting from general low learning, possibly due to using a single-session training. In this longer training study, while 7-8 year-olds still showed low learning at post-test, 11-12 year-olds *did* show generalisation in the Discrimination, minimal pair Picture Identification and Vocabulary tasks, meaning there was opportunity to see a possible variability effect. Nevertheless, none was found.

One possible explanation is that the task demands were too high for learners to properly benefit from having HV input, as increased variability is known to make a task more complex. Recall that Sinkevičiūtė et al. (2019) did not find a variability advantage for vocabulary learning in either 11-12 year-olds or 7-8 year-olds, although they did find it in adults. They attributed this to the greater difficulty of the multiple-talker input in training. This is also seen in training in the current study, at least with 11-12 year-olds who do better with LV than HV training. Moreover, the current task did not just have talker variability that learners had to adjust to: it also trained the learners on four different vowel contrasts at the same time. Most studies using the HVPT paradigm train learners on one phoneme contrast at a time; there are some that go beyond this (see Section 1.3.2), and this may be beneficial for adult learners (Nishi & Kewley-Port, 2007), but no studies have previously done this with child participants. It is possible that this added complexity, in combination with increased talker variability, might have been too cognitively demanding for the children to be able to fully benefit from the advantages that have been thought to result from increasing input variability. I return to consider possible reasons for not seeing a benefit of high variability and how this fits with previous literature in the General Discussion in Chapter 5. Finally note here that, although this study finds no evidence for an HV benefit, it also did not find an LV benefit, failing to replicate Giannakopoulou et al. (2017), who found such a benefit in 7-year-olds in a discrimination task. Again, since 7-8 year-olds show no generalisation in this study, a lack of difference between

conditions is hard to interpret. For the 11-12 year-olds, it is unclear whether an LV effect should be predicted, however note that, in all of the tests where they showed pre- to post-test improvement (Discrimination, Picture Identification with minimal pairs, Vocabulary) the evidence for an LV benefit was ambiguous. This makes it hard to draw firm conclusions about the role of variability from the current data. It is possible that, with a larger sample an LV benefit might have been seen.

4.4.3 Conclusion

To conclude, the current study trained native Dutch 7-8 year-olds and 11-12 year-olds on four SSBE vowel contrasts using a two-week phonetic training paradigm that used both meaningful pictures and familiar orthography in training. While both age groups showed improvement during training, the 7-8 year-olds were not able to generalise this acquired knowledge to any of the post-test tasks using novel speakers. 11-12 year olds, on the other hand, were able to generalise beyond the trained materials in most tasks. Crucially, as in Chapter 3, this study did not find any effect of variability in any of the tasks.

Chapter 5. General Discussion

This thesis explored the effect of high and low variability input in phonetic training for non-native speech sound learning in children. This aim was explored through means of two experimental studies: Study 1, consisting of 3 separate single session experiments with 7-8 year-olds and adults, and Study 2, which was a large scale 8-session training study conducted with 7-8 year-olds and 11-12 year-olds.

The single session experiments in Study 1 explored the HVPT paradigm training English speaking adults and 7-8 year-old children on Dutch diphthongal vowel contrasts. Adults and children were both naïve learners of the target language and were thus matched in their L2 exposure. The first experiment used an identification task involving orthography-like symbols. However, this showed that while adults showed substantial evidence of learning, children's performance was much worse, with ambiguous evidence as to whether they were above chance even in training, and substantial evidence that they were not in generalisation. Since learning was low, the focus of the study shifted to exploring whether the form of the training task was leading to low learning and generalisation in children through means of a set of follow-up experiments. These experiments found that a learning task where participants mapped whole words to meaningful pictures led to greater learning compared with the orthography-like training task, even given the same auditory stimuli and amounts of exposure. Interestingly, across all experiments, even where participants showed evidence of generalisation to new

voices in the post-test identification task (which was the case for adults in two out of three experiments, and for children in one experiment), they did not show any evidence of improvement on a pre/post discrimination task. In addition, returning to the initial hypotheses, even where participants showed learning, there was no difference between the HV and LV variability conditions, with evidence for the null for an HV advantage in generalisation for adults and children in Experiment 2, and evidence for the null for adults in Experiment 1.

After considering that the relatively low levels of learning in Study 1, particularly in discrimination, and the lack of HV benefit might be due to the use of a single session training study, Study 2 moved to use a longer, multi-session training paradigm. Anticipating greater difficulties in recruitment for a study that required a large time investment from schools, native Dutch learners of English were chosen as the participant group. There is a strong interest in English education in primary school in the Netherlands, as evidenced by ongoing pilots investigating the effect of teaching English from age 4 (EP Nuffic, n.d.). This interest provides primary schools with a strong motivation to participate in research involving English, facilitating participant recruitment. Although the original plan was to recruit adults, pilot testing suggested that this would lead to ceiling effects, and thus 7-8 year-olds and 11-12 year-olds were recruited. Both groups participated in a full-length two-week training study, in which they were trained on English vowel contrasts. Given the results of Study 1, a picture-based training paradigm was used to teach the Dutch children English vowels, but in this case orthography was also

included in training, with tests included which could assess separately whether pictures and/or orthography were used in learning. Unlike Study 1, all tests were included both pre- and post-training, and a production test and vocabulary learning test were also included. The key questions concerned whether training would result in improvement on pre/post-test tasks, whether variability in training input affected performance, and whether these factors differently affected the two age groups. Overall, and contrary to the notion of greater plasticity in younger learners, older children not only outperformed younger children at pre-test, they also improved more across Training, and showed more improvement as a result of training in the Discrimination, minimal pair Picture Identification and Vocabulary tasks. In fact, although younger children showed improvements during Training itself, they made no improvement at all on any of the pre/post-tests, suggesting that they did not generalise their learning to untrained talkers. Interestingly, even the 11-12 year-olds, who did improve on most tasks, did not show improvement on the task which tested whether they were able to make use of the orthography during Training, with evidence of the null. However, in contrast to Study 1, 11-12 year-olds *did* improve their discrimination ability as a result of training. Critically, there was no evidence of a HV benefit anywhere (i.e. even for the 11-12 year-olds, who did show learning), with evidence for the null. There was also no evidence of an alternative LV effect (following Giannakopoulou et al. (2017)); in fact, the only task that showed any variability effect was the training task itself, where there was an LV benefit for 11-12 year-olds, although this was ambiguous for 7-8 year-olds.

The rest of this chapter will consider the implications of these findings. The discussion is organized into a discussion of the role of orthography versus pictures in training, a brief discussion on discrimination performance in these experiments, and the findings with regard to age differences, before discussing the key concern of the thesis: the consistent yet surprising lack of variability effects. Finally, the methodological contributions of this thesis will be discussed, before suggesting future directions and concluding the thesis.

5.1 The role of orthography vs pictures

One key finding of this thesis relates to the role of orthography in L2 phonetic training. In Study 1, children had difficulty in learning non-native vowels by mapping to an orthography-like system, while a meaningful picture-based system led to much greater learning, and even adults showed stronger learning with the picture-based task. In Study 2, while for 11-12 year-olds picture-based phonetic training proved successful, and resulted in learning that generalised to other tasks, an orthography identification task did not show any improvement as a result of training.

In both studies, providing participants with orthography which directly represents the non-native phonemes does *not* seem to help them learn the novel phoneme categories. However, note that the 'orthography' (i.e. orthography-like symbols) used in Study 1 differs substantially from the English orthography used in Study 2,

and therefore, the reasons why participants did not use orthography in training might differ in these two cases. In Study 1, participants might have found it easier to map the stimuli to meaningful pictures in Experiment 2, because the stimulus can be processed as a whole, while for mapping stimuli to orthography-like symbols in Experiment 1 and Experiment 3, an abstraction process needs to occur first. Rather than processing the whole stimulus, a participant needs to break the stimulus down into its phonemes, and then needs to discover the consistent pattern in the mapping of these phonemes to the symbols. If the phonemes the learner is presented with are then additionally confusable non-native sounds, discovering the underlying pattern might be quite a lot harder, if not impossible to achieve. This could explain why children and even adults found this task hard. In Study 2, participants did not have to go through this abstraction process, as actual English orthography is provided. The children can use their knowledge of Dutch orthography (as well as any initial knowledge of English orthography they might already have) to tell them which are the vowels and which are the consonants. This means that having to work out which part of the word the vowel links to should not be a problem in this case, and thus cannot explain the lack of improvement in the orthography task of Study 2.

Nevertheless, children still did not seem to use the orthography in Study 2. This is quite unexpected. Theoretically, orthography might be expected to be helpful, as it helps identify possible ways of grouping variable spoken data: every instance of a phoneme is acoustically slightly different, so having a stable orthographic cue

could help learners identify that, despite these differences, the instances belong to the same category. It is thus quite interesting that children did not seem to use the orthography to help create these categories, as the phonetic training literature suggests they might benefit from it (see Section 1.3.1). As it was not possible to test adults for this study, it is unclear whether they would make use of the orthographic cues. There is an additional literature exploring the role of orthography on spoken word acquisition, though again results here are mixed. Simon et al. (2010) showed that adding orthography resulted in better performance during training, but did not affect test performance; similarly, Escudero (2015) found adding orthography mostly has no effect on spoken word learning, although there is some positive effect specifically for easier contrasts. Bürki, Welby, Clément, & Spinelli (2019) showed adding orthography positively influences word learning, while Showalter & Hayes-Harb (2015) found negative effects for word learning when the orthography is completely unfamiliar. Adding orthography has also been found to affect the pronunciation of L2 words, though again the direction of the effect is more mixed: Nimz & Khattab (2019) found mostly beneficial effects of adding orthography, while Hayes-Harb, Brown, & Smith (2018) found a negative influence on production.

The finding that children do not seem to use orthography also goes against the benefit of adding orthography for children's vocabulary learning found in Ricketts et al. (2009): they taught 8-9 year-olds novel spoken word forms and trained them to associate these with pictures of novel objects, or with pictures and orthography

together (although children were not explicitly instructed to make use of the orthography). Children showed evidence of having learnt the orthography they were presented with, and showed stronger learning on the words presented with both pictures and orthography compared to those where pictures only were presented. These results might suggest that orthography has a beneficial effect on children's word learning in their L1; however, this does not seem to extend to phonetic training in the L2 where the inclusion of orthography proves too hard for children. In part, this might be because children at the ages tested in this thesis are still in the process of acquiring their L1. The ability to consistently categorise sounds in children's L1 has been shown to continue at least up to the age of 12 (Hazan & Barrett, 2000), and simultaneously children are in the process of mastering the writing system in their native language (see Treiman & Kessler (2014) for an in-depth overview of this process). Additionally, they might still be developing their phonological awareness, directly influencing their ability to learn the phonological structure of spoken words (Gathercole, 1999). This explanation is in line with the account in Shinohara (2014) discussed in Section 1.3.4, which explained the poor learning with their 6-8 year old participants in terms of a lack of phonological awareness. Providing children whose L1 is still very much in development with L2 stimuli that are supposed to be mapped to an orthography-like system of unfamiliar symbols might at that age just be beyond their cognitive capacities. Recall that there are some studies in the phonetic training literature that did use orthography-based training with children. However, all of these used

orthography where there was some kind of iconic relationship between the symbols and the phonemes they represented: in Giannakopoulou et al. (2013) the length of the digraph <ee> versus monograph <i> corresponded to the vowel length, and in Heeren & Schouten (2010) the labels 'long t' and 'short t' corresponded to the length of the closure. It thus seems that mapping phonemes to non-arbitrary symbols is what is helpful in this case.

Compared with using arbitrary representations of phoneme categories, children seem to learn better by mapping *whole words*, including the non-native vowels, to something conceptual. Perhaps this is the case because it reflects how such a situation might play out in real life: people primarily use language to communicate, not to map individual phonemes to symbols without contextual cues. This has implications for L2 speech research with children, where researchers might want to consider the ecological validity of the situation in which their stimuli are presented: a more 'unnatural task', further removed from real-life language learning, might not be very effective in eliciting any learning at all.

Of course children do manage to learn the orthography of a second language. This raises the question of whether it is better to learn a spoken language first and then learn orthography once the phonological system is well-established, or whether it is reasonable to learn both at the same time, and that children would eventually pick up on the orthography as well. In terms of the phonetic training paradigm, if using orthography is difficulty, it might be necessary to explicitly train orthography on its own, without pictures, for children to start using this cue; it is

possible that they rely on the easier cue, i.e. pictures, where they can, but could make use of the orthography if they needed to in the absence of pictures. Note that from Study 2, it is impossible to say whether orthography *hurts* learning, i.e. whether performance would have been better without orthography in the training task, or whether it is just ignored. Suggestions for how to further investigate the role of orthography in L2 speech acquisition are discussed in Section 5.6, where future directions will be discussed.

5.2 Discrimination performance

An interesting finding of this thesis concerns the effect of training on participants' discrimination abilities. While overall performance was well above chance, Study 1 showed that adults and children did not change their discrimination abilities as a result of a one-session identification training, with evidence for the null for an improvement from pre- to post-test even in conditions where participants otherwise showed learning and generalisation. However, in Study 2, while the younger children showed no improvement in Discrimination as a result of Training, the 11-12 year-olds did show improvement on the Discrimination task as a result of two weeks' worth of phonetic training. Importantly, the younger children did not show this improvement in any of the pre/post-test tasks, suggesting their lack of improvement is not specific to their discrimination abilities.

It is in principle possible that this difference in improvement on the discrimination task is to do with the age of the participants, i.e. there might be something special about children aged 11-12 that allows them to improve in discrimination while younger learners and adults do not improve on the same task. However, note that this does not fit with the literature, since participants of the same ages as in Study 1 have been found to improve in discrimination (see (Iverson et al., 2012) for adults, and Evans & Martín-Alvarez (2016); Giannakopoulou et al. (2010) for 7-year-olds). Crucially, the difference between this literature and the current Study 1 is the length of training.

This leaves two possible ways to explain this difference across the studies. Firstly, the issue that prevented an improvement in discrimination abilities in Study 1 might have been the training length. Earlier studies such as Heeren & Schouten (2010) have found a similar lack of change in discrimination after a short amount of training, and have suggested that discrimination abilities might only change after extended exposure, as this is needed to adjust the phoneme representations. This is in line with predictions made by Flege's Speech Learning Model, which predicts non-native contrasts only dissimilate from having been mapped onto a single native phoneme through extended exposure.

Another possible explanation for the observed changes in discrimination is that unlike the participants in Study 1, the 11-12 year-olds in Study 2 are not naïve learners. It might be the case that naïve learners find it harder to change their discrimination abilities compared to learners who have had some exposure to the

L2. However, the literature does not support this: while relatively few studies use naïve learners in phonetic training, improvement in discrimination has been found in all that do (Dong et al., 2019; Fuhrmeister & Myers, 2017; Kartushina et al., 2015; Sadakata & McQueen, 2013).

An interesting final note on discrimination performance is that generally, despite only showing improvement in one age group in Study 2, performance on the discrimination task was actually quite high throughout the thesis, especially compared to performance on the various identification tasks. This is in line with predictions from Best's Perceptual Assimilation Model (described in section 1.2), where discrimination performance is predicted to be better than identification performance when two phonemes are assimilated into a single category with a category goodness difference (e.g. SSBE /e/ might be a better fit to Dutch /ɛ/ than SSBE /æ/).

5.3 Age differences

This thesis consistently found a difference in performance between older and younger learners. Older learners did not only show better overall performance, including performance at pre-test, but they also showed better learning than younger learners in both studies. In Study 1, there was substantial evidence that adults had a steeper learning slope during training than 7-8 year-olds in both Experiment 1, as well as for the Lab-based adults from Experiment 2. In Study 2

there was substantial evidence of greater learning in 11-12 year-olds than in 7-8 year-olds in all of the perceptual pre/post-tests except the orthography task and the non-pairs of the Identification task (where no improvement was found for either age group, for reasons discussed above). Why might it be the case that older learners consistently outperform younger learners?

The fact that older learners were better at pre-test is not surprising, as most of the literature has found this (see Section 1.3.4). However, in the majority of studies that find this result, this could be explained by older learners having previously had increased exposure to and more experience with the L2. While this is true in the current Study 2, it is not a possible explanation in Study 1, where all age groups were naïve learners. The only study to compare age groups with learners who were all naïve to the L2 is Wang & Kuhl (2003), where American English adults and children learned Mandarin tones. Their explanations for the age difference in performance was that either children did not have the cognitive maturity to be able to perform the task as well as adults did, or that the suprasegmental tones were so non-nativelike that they could not be mapped to an existing L1 category for either age group. However, since the current thesis also found evidence for greater improvement in adults over children for Study 1, the latter explanation proffered by Wang & Kuhl (2003) does not hold: in the current study segmental phoneme categories are used that are relatively close to existing native phoneme categories. This leaves the explanation that due to a difference in cognitive maturity, adults found the task easier than children in both Wang & Kuhl's task as well as the

current Study 1. This suggests that the fact that adults find the task easier can also lead to them gaining more from this type of training.

The current results fit most clearly with the findings of Shinohara (2014), where younger children (aged 6-8) performed worse than older children (aged 8-12), although older children did improve more than adults. They explain this difference between younger and older children through their phonemic awareness skills: younger children are still developing their phonemic awareness, which is needed to learn to map orthography to sounds; older children have learnt to do this, and thus outperform the younger children. However, while this could explain the results of Experiment 1 and 3 of Study 1 where an orthography-like system was used in Training, this is not a possible explanation for the results of Study 2, where orthography is not the only cue in training nor the only factor tested in the pre/post-tasks. The performance difference between 7-8 year-olds and 11-12 year-olds there thus needs further investigation.

The current data is also consistent with findings from Giannakopoulou et al. (2017); Heeren & Schouten (2010) and Wang & Kuhl (2003) in that it does not find evidence for greater amounts of improvement in younger learners versus older learners. Note that although these studies also did not find a benefit of older learners improving more than younger ones, they used frequentist statistics which means it is impossible to differentiate between ambiguous evidence and evidence for the null for either of the two hypotheses. This makes it hard to tell whether there is no difference between the amounts of improvement in either age group,

or whether the data they collected is insensitive and more data would show a clear effect.

However, there are two studies that find a different result: 7-8 year-olds showing greater learning than adults in Giannakopoulou et al. (2013), and older children (8-12) and adolescents (15-18) showing greater learning than adults in Shinohara (2014). First, in Giannakopoulou et al. (2013) children showed greater learning than adults in both a perceptual identification task as well as an AXB discrimination task. For the discrimination task, this is possibly explained by the fact that adults' performance at pre-test was already near ceiling and thus left little room for improvement while children started out lower with more room to learn. However, this cannot explain the perceptual identification data, where adults and children started at similar levels at pre-test. Then, Shinohara (2014) found greater learning in 8-12 year-olds and 15-18 year-olds than in adults in identification as well as discrimination tasks. Like Giannakopoulou et al. (2013), Shinohara (2014) explains this difference between younger and older learners as there being a plasticity benefit for younger learners, although he notes this only holds provided the children are old enough to have sufficiently developed their phonological awareness so that they could benefit from the task. As noted above, the current studies do not rely on the use of orthography, so the explanation provided by Shinohara (2014) in terms of the role of phonological awareness in mapping phonemes to orthography does not apply. Still, it is possible that more generally younger children's lesser cognitive abilities impede them in the current task, as

discussed above, and this might outweigh any plasticity effect. Since the current thesis did not compare the performance of adults and 11-12 year-olds, it is not possible to tell if 11-12 year-olds would show a plasticity effect over adults in either of the studies.

What does this mean for the critical period hypothesis discussed in the introduction to Chapter 1? Recall that this hypothesis was based on naturalistic data, where children learn language through immersion. In naturalistic data, children have been shown to display a better phonological acquisition of the L2 if they started learning at a younger age (see the discussion at the start of Chapter 1). Assuming that in this context the idea of 'the younger the better' does hold for children as young as 7-8 years old (and the review by Ioup (2008) would suggest it does), there must be something about the type of input that learners are exposed to in more naturalistic contexts that is somehow better for children and allows them to benefit more than adults do, compared with the type of input used in training paradigms like the one used in this thesis. One obvious difference is that there is just so much more data available in naturalistic contexts, as learners hear the language constantly rather than in around ten sessions of maximally half an hour each. Additionally, hearing language out of context in the training paradigm might be less beneficial than hearing it used in a communicative context in naturalistic language learning situations, where the context can provide additional cues to the meaning and thereby to potentially confusing contrastive phonemes. It might also be the case that naturalistic contexts require a different part of the

phonological awareness skills that have developed at an earlier age than the phonemic awareness skills required for the current type of training. Perhaps children's plasticity benefit, therefore, only shows in naturalistic circumstances.

However, a final possibility is that there really is no plasticity benefit, and that children outperform adults in real language learning situations simply because they get more input than adults overall because they go to school. This explanation is tentatively supported by data from Hartshorne et al. (2018), although this study looked at syntactic rather than phonological learning. In this study the L2 learning rate was found to decline around the age of 17, coinciding with the age around which people tend to leave full-time education. Adult learners, while equally immersed in the L2 setting, might receive less (or less rich) language input in their daily life compared to the amount of exposure children receive in schools. Unfortunately, it is difficult to untangle the role of exposure from the role of age of acquisition in terms of naturalistic L2 acquisition settings. Studies like those performed in the current thesis potentially allow us to pull apart the role of age and amount of exposure, but it remains to be seen what type of training is needed to allow plasticity benefits to be seen.

5.4 Variability effect

Finally, the most striking finding of this thesis is the consistent lack of a variability effect in learning, either in adults (Chapter 3), 11-12 year-olds (Chapter 4) or 7-8

year-olds (Chapter 3 and Chapter 4). The only task that showed any difference between the variability conditions is Training itself. Across all the experiments presented in this thesis, there was generally numerically greater performance in the LV than HV condition in training, although the evidence for this was only substantial with lab-based adults in Study 1 Experiment 2, and 11-12 year-olds in Study 2. This difference during training is in line with the results from L2 vocabulary learning in children reported in Sinkevičiūtė et al. (2019), where although LV input outperformed HV in training for 7-year-olds, no evidence of a variability effect was found, with evidence for the null (since this study used Bayes factors). A similar variability effect was found in the current online adults of Study 1, Experiment 2, where those in LV training improved more across blocks than those in HV training. This lack of a variability effect on generalisation is an unexpected result given the literature reviewed in Section 1.3.5.

On the other hand, the number of studies which make a direct comparison between HV versus LV training input is actually relatively small – just three studies compare input variability for adults (Lively et al., 1993; Sadakata & McQueen, 2013; J. W. S. Wong, 2012, 2014), at least for learning phonemes, as opposed to tones, where a benefit is linked to aptitude if it is found at all (Dong et al., 2019; Perrachione et al., 2011; Sadakata & McQueen, 2014)). There is also the study by Evans & Martín-Alvarez (2016) with children, however, this only found the HV benefit via an interaction with item novelty, which differs somewhat from the other studies. However, there is some related evidence from studies of voice

identification (Lavan et al., 2019) and accent identification (Clopper & Pisoni, 2004). There are also more studies in the field of vocabulary learning finding HV input to be beneficial (e.g. Antoniou, Wong, & Wang (2015); Barcroft & Sommers (2005, 2014); Sinkevičiūtė et al. (2019); Sommers & Barcroft (2006, 2011)), and despite having vocabulary tests as part of Study 2, no variability effect was found in the current thesis.

As discussed in the introduction in Section 1.5, one potential explanation for these differences with the published literature is publication bias, i.e. the fact that it has traditionally been difficult to publish null effects. This means that there are likely studies which have been conducted investigating the role of variability in phonetic training input, and which have not found an effect but which are not published. Given this, a replication of the original (e.g. Lively et al. (1993); Logan et al. (1991)) might be warranted where high versus low variability input are directly compared in phonetic training with adults.

The problem of publication bias will be considered in more detail in Section 5.5 on the methodological contribution of this thesis. For now, note that there are good theoretical reasons why training with multiple talkers should help generalisation to novel talkers: it should be easier to generalise to a novel speaker after having heard multiple examples of ways in which a speaker might vary than after having only heard one specific speaker. However, it may be the case that there are circumstances under which this benefit is not seen. What circumstances might prevent an HV benefit from appearing in the current study?

One general reason for not finding a HV benefit would be if the benefits for generalisation are outweighed by the increased complexity that comes with adding variability to the input. This is in line with the explanation given in Giannakopoulou et al. (2017); Sinkevičiūtė et al. (2019). Recall that multiple-talker input has been found to be harder to process in certain tasks, even in an L1, as it takes up more working memory capacity (Martin et al. (1989); Mullennix et al. (1989); Sommers & Barcroft (2006), see Section 1.1.2 for a detailed discussion). Since variability can thus be harder to process, this means it would be possible to not always see a variability benefit in generalisation. This would particularly be the case if some aspects of the tasks themselves are also hard for the particular learners. The following paragraphs will consider this in the context of the different age groups and tasks in each of the studies of this thesis.

In Study 1, the adults found the shape-based training task surprisingly hard. However, they did not find the picture-based training task hard, so there is a possibility of finding variability effects there. One reason that those effects were not found, however, could be the rather short length of training, meaning it takes more time for an HV benefit to emerge. For the 7-8 year-olds in Study 1, working memory limitations could make HV input generally harder. They too found the shape-based training task hard, but did reasonably well with the picture-based training task; however again, although an HV benefit might thus be expected to show up here, it was not found. The reason it does not show could be the training

length, as for adults, but it could also be a combination of their age and cognitive capabilities to perform the tasks.

In Study 2, the training is lengthened which could potentially help untangle what might cause the lack of HV benefit in the 7-8 year-olds of Study 1. However, in addition to a longer training, another addition was made, namely to test more vowel contrasts: rather than one three-way contrast in Study 1, Study 2 tested four two-way vowel contrasts. This made the task significantly harder, and 7-8 year-olds showed floor effects in the pre/post-test tasks, so again perhaps this added complexity did work against an HV benefit in this case. The 11-12 year-olds in Study 2 did not have quite as many difficulties with the tasks as the 7-8 year-olds, evidenced by the lack of floor effects and their showing evidence of learning. An HV benefit could thus be expected, but once again did not show up. It is possible that perhaps the task was still too hard to fully show a variability effect; recall that in Sinkevičiūtė et al. (2019) an HV benefit was found in adults but not in either 7 or 11-year-old children. Further exposure might have been needed to elicit an HV benefit, so a longer training session might still be required to find the coveted variability effect.

All in all, it is thus possible that HV benefits could have emerged, at least with adults and possibly also in 7-8 year-olds, if Study 1 had had a longer training period. Moreover, it is also possible that an HV benefit might have emerged in the 11-12 year-olds and possibly with the 7-8 year-olds if the training had been even longer in Study 2, and/or if only a single vowel contrast had been trained.

A final possibility that could explain the lack of HV benefit in all age groups, is that whether a variability benefit emerges could depend on idiosyncratic details of the items used in training and at test. It could be the case that previous literature where an HV was found actually showed such a benefit because one of the voices used in HV training happened to be similar to the voice(s) used at test. Note that not all of the literature fully counterbalances the voices used in training and test (as was done in both studies in the current thesis), meaning that there could be talker-specific properties that are causing an HV benefit. In some sense this is a problem with those studies. At the same time, it is true that the more voices are used in training, the higher the chances of finding similarities between trained and test voices. However, in order to fully cover the 'space' of possible types of voices, a very large number of voices would need to be trained; the fairly common choice of using four different voices in training would only begin to cover this.

Although no HV benefit was seen in the current data, with evidence of the null in some cases, the current thesis also did not find evidence for the alternative hypothesis that LV input would be more beneficial for learning and generalisation. The alternative prediction of an LV benefit in children was based on Giannakopoulou et al. (2017) who found no variability difference in adults but found that children benefitted from LV training input even in the generalisation tasks. It is worth noting, however, that the authors of the paper were very cautious in interpreting their results, especially given that the LV group - by chance – started at a lower performance level than the HV group, leaving more room for

improvement throughout (although the *glmer* analyses should have controlled for this to some extent, it is possible it nevertheless influenced results). Another study which found an LV benefit was Evans & Martín-Alvarez (2016), who found this benefit specifically in production but not in perception. Given this information, it is quite possible that the LV benefit found in both of these previous studies was actually a false positive (type-I error). However, it is important to note that in the current thesis, the evidence for an LV benefit was generally ambiguous. In fact, there was nowhere in the data where there was evidence of overall learning, yet a null effect for the LV benefit (unlike for the HV benefit). Ambiguous evidence means that the possibility that an LV benefit might be seen with a larger sample cannot be ruled out. Following the line of reasoning developed in Giannakopoulou et al. (2017); Sinkevičiūtė et al. (2019) discussed above, such a benefit would indicate that any benefits of HV input for generalisation are fully outweighed by ease of processing single-talker input.

In conclusion, the current thesis contributes to the field of phonetic training by providing evidence for the null for finding an HV benefit for adults and 7-8 year-old English naïve learners of the Dutch /ɑu/-/ø:/-/œy/, as well as for 7-8 and 11-12 year-old Dutch learners of SSBE GOOSE-FOOT /uː/-/ʊ/, STRUT-LOT /ʌ/-/ɒ/, DRESS-TRAP /e/-/æ/, and FLEECE-THOUGHT /iː/-/ɔː/. No LV benefit was found either, however here the evidence was ambiguous. The discussion above suggests that whether an HV or LV benefit is obtained likely depends on the difficulty of

the training materials and tests for the learners in questions. It is for future work to try and tease apart which types of stimuli and training do and do not lead to an HV (or LV) benefit, and why they do so.

5.5 Methodological contribution

Recall that, as noted in Section 5.4 above and discussed at length in Section 1.5 of the introduction, there has been a problem with publication bias in the field of phonetic training, as well as in psychology more broadly, so that studies with null results are not reported. Where studies with null results *are* reported, null effects are often over-interpreted as showing evidence for the null, when actually statistics usually do not provide this evidence, meaning the result could be a type-II error.

A major contribution of the current work is that, in both of the studies of the thesis, Bayes Factors are used to quantify evidence for the null, as well as H1. Only two studies in the literature have used Bayes Factors to investigate a variability effect before (Dong et al., 2019; Sinkevičiūtė et al., 2019). These studies, however, do not use Bayes Factors throughout, but only use them to follow up on null results. This thesis contributes by being the first to use Bayes Factors throughout, following Dienes & McLatchie's (2018) advice to provide 'a B for every p'. One difficult aspect of this is estimating the size of the predicted effects in order to inform H1. The present thesis contributes by demonstrating how this can be done even when there

is no prior study with sufficiently similar materials on which to base the estimation.

Note that using Bayes Factors also has the advantage of providing a continuous measure of evidence (unlike p-values). Experimenters can, therefore, continue to add to the sample in order to obtain stronger evidence (i.e. there is no optional stopping problem (Dienes, 2016); although this has been disputed (Grünwald & de Heide, 2018), the process only appears to be a problem when default, unjustified priors are used (Rouder, 2019)). This means that in principle, it would be possible to continue collecting data wherever the Bayes Factors suggest that the data is ambiguous, although practicalities meant that this was not possible for the current thesis. This could offer an enormous benefit for research with populations that are difficult to recruit, so that data is not 'wasted' when an experiment comes back as having ambiguous results.

Finally, this thesis contributes by pre-registering one of its studies. Preregistration is one of the major proposals for overcoming publication bias in the literature. Overall, it was not just useful to do good science, but preregistration was also helpful in producing the research itself. The process of preregistering meant it was necessary to have clear hypotheses before data collection had even begun, which was helpful in clarifying the experimental design as well as in thinking about possible analyses. Still, it did not prove possible to follow the preregistered plan to the letter (which can perhaps be expected when working with children). For example, tasks could not be performed exactly as planned (e.g. the category

boundary task could not be administered at all, and only a subset of children partook in the production task of Study 2), and analyses were planned to be based on effects that were not found (e.g. the preregistration planned to base estimates for interactions with input variability on the main effect of session, but there were cases where participants did not improve at all from pre- to post-test). There were also some cases where the preregistration contained errors, or where a greater understanding had since been gained regarding the appropriate analyses at the point at which these were performed. Although this thesis could have followed exactly what was laid out in the preregistration, it was deemed much more important to do analyses that were appropriate. Therefore, a very transparent approach was adopted in which any changes that were made to the preregistered analyses were clearly indicated (see Appendix III). Note also that preregistration allows for additional exploratory analyses. In the current thesis, additional analyses could for example have looked at performance differences between the vowel contrasts, or at performance in response time data. Overall, preregistration was a very useful addition to the research process in helping to prevent bias and also to clarify the design and analyses upfront. However, an important lesson was that the preregistration should not constrain what analyses are performed, and that rather than sticking to preconceived notions of analyses that end up being wrong or inappropriate, transparency will always lead to better science.

5.6 Limitations and implications

The studies in this thesis contributed to our understanding of phonetic training and the role of variability. However, there were several limitations that should be considered when contemplating the implications of this thesis, many of them touched upon in previous sections. Most of the limitations of Study 1 were addressed in the implementation for Study 2, so this section will mainly consider limitations to the latter study, before considering the implications this has for the phonetic training literature more generally.

Firstly, the question remains whether all tasks in the pre/post-test battery of Study 2 were truly appropriate for the youngest age group. There are signs that this might not have been the case for some of the tasks, in particular for the Vocabulary and Orthography Identification task, which means this might have affected the results of the younger age group. This also suggests that these particular tasks might have been harder for the 11-12 year-olds as well compared to the other tasks in the test battery; it is unclear how this may have affected their overall results, and further investigation might be warranted.

In addition to task-specific effects on participants' performance, there might also have been an effect of orthography more generally. As discussed in Section 4.4.1, English orthography is not entirely consistent, and having an orthographic cue that might have been less reliable for vowel perception than what they were used to from Dutch might have been detrimental for the children trying to use it to learn

the non-native contrast. Moreover, it could be the case that having orthography in general would have been distracting for the beginning readers, as it takes up resources to spend on a potentially less informative cue rather than using them to learn from the cues that they can use more easily. Additionally, there might have been transfer effects from Dutch orthography that influenced performance on the various tasks, though the extent to which this might have happened and what kind of effects this will have had on the overall performance cannot be determined from the current tasks alone.

Furthermore, there is the fact that vocabulary knowledge might have affected overall performance throughout training and at pre/post-test. Since none of the participants in Study 2 were entirely naïve learners, they might have been more or less familiar with the particular vocabulary used in the training and tests before the sessions took place. This may have affected their performance, where those children who were already familiar with the word meanings could fully focus on learning to distinguish the minimal pairs on the basis of the non-native vowel contrast, while those who were less familiar with the meaning would have had to focus on the word meaning as well in order to be able to map the words to the pictures.

It is important to remember that the Production results as they currently stand can only have a very limited scope due to the technical issues in recording and the loss of participants. Moreover, the production rating has only been done by one speaker while a larger team of raters would be preferable, especially given the issues this

particular rater had with one of the vowels. A more substantial group of raters might have been able to uncover more nuanced changes in the pronunciation. The production data has also not been analysed acoustically for this thesis, but it is entirely possible that the children are starting to produce a change in their vowels that is not picked up on in the native speaker ratings. Further acoustic analyses would be able to provide more insight into the possibility of change being underway that is acoustically detectable but not yet categorically classified as such by native speakers.

In spite of these methodological issues, this thesis has implications for the phonetic training field as a whole. Firstly, this thesis shows that an adapted version of the phonetic training paradigm can successfully be used with naïve learners of an L2, and in particular with child learners as well. This version of the paradigm also turned out to be suitable of beginning but not naïve child L2 learners, though the poorer results of the younger learners showed the importance of ensuring the tasks used in these types of paradigms are suitable for the specific age group in terms of their task demands. This thesis also showed that the role of input variability in phonetic training might not be the same for children as it has been shown to be for adults: this thesis found no evidence for a variability benefit, suggesting participant age groups and task demands might play an important role in this result. This ties in with work from (Ingvalson et al., 2013; Perrachione et al., 2011; Sadakata & McQueen, 2014) who showed that learners' aptitude, in this case for learning Mandarin tones, was strongly predictive of whether they benefitted from

receiving high variability input; note that no such link between a variability benefit and individual aptitude was found in (Dong et al., 2019). The extent to which the key finding from this thesis transfers to other participant groups and task types is thus something that requires further investigation still.

Finally, it is important to return to the question of what we mean by phonetic learning, in light of the findings of this thesis. Note that it is impossible to draw conclusions from those experiments in Study 1 where no improvement as a result of training was found. However, for the cases where learning as a result of training was found in both Study 1 and Study 2, the results did not find improvement across all tasks: tasks relying on identification ability were often the first to show improvement, while those relying on discrimination skills and production showed little to no improvement. While unfortunately the category boundary task in Study 2 malfunctioned and can thus not shed any light on whether there was clear evidence for change to the underlying categories, the diverse pattern of results seen across different tasks does suggest that whatever phonetic learning did take place most likely occurred at a surface level rather than at the level of the phonetic representations, in line with results from (Heeren & Schouten, 2010; Iverson & Evans, 2009). If the underlying representations were undergoing change as a result of training as well, more uniform patterns of improvement might have been expected across the different tasks in this thesis. However, this suggested surface-level phonetic learning does not mean that phonetic training itself is any less useful: improved phoneme perception and identification abilities are key in day-

to-day interaction, so ultimately phonetic training can still prove to be very helpful in improving learners' ability to function in an L2 environment. Further research is needed to truly establish the extent to which phonetic training affects the phonetic representations, and the different types of phonetic learning it might initiate.

5.7 Future directions

The results from this thesis have implications for future research, as well as more practical implications. It is very important to keep in mind that the results of this thesis are limited to Dutch learners of English and English learners of Dutch, trained on a specific set of vowel contrasts, and that caution should be exercised in extending them to other groups of learners or other training regimes, as more research would be required.

Nevertheless, there are some potential practical implications of the finding that multiple talker input does not necessarily boost learning and generalisation. For companies developing training software that teaches learners particular non-native contrasts, this result suggests that it might not be necessary to include a plethora of voices, at least initially. Moreover, translating this to a classroom setting, teachers are likely better off focussing on providing the learners with as much actual content and exposure as possible, rather than solely focussing on the number of talkers used in this exposure.

Another important question for both developers of language training materials and teachers is whether it is better to introduce orthography in the early stages of learning a new language. The results of the current study suggest that orthography representations might not be helpful for child learners. However, there is more to be done to fully understand the result. First, while Study 2 showed no evidence of orthography in itself being learnt, it might still be beneficial for learning to have redundant cues in the form of both orthography as well as pictures. One way to test this, would be to repeat the study with the same training but with an additional pre/post identification test where both pictures and orthography are presented at once (again using a novel voice to test generalisation), and seeing how performance on this task compares to performance on the orthography-only and picture-only identification tasks. However, it is also possible that adding orthography might actually *harm* learning. This question could be investigated further by running the same study as was run in Study 2 here, but instead of using both pictures and orthography in training, using pictures only and seeing if children do better on the post-tests improvement for that version than they did in the current study. This would shed more light on the exact role of orthography in phonetic training.

Another question raised by this thesis relates to performance differences as a result of age. There is a widespread assumption that younger learners will outperform older learners, leading to beginning language teaching at a younger age (see for instance EP Nuffic (2015) for an example in the Netherlands, but note that Goriot (2019) showed starting at an earlier age made little difference to actual L2

competence of this same cohort). This thesis adds to the current research in showing that younger learners do not necessarily outperform older learners, and that multiple factors beyond just age might play an important role in the overall learning performance. Further research is needed to fully establish if and when a benefit for younger learners might emerge. For example, in the case of phonetic training, a benefit for younger learners might only emerge with multiple session training. This would explain why it was not seen with a single training session in Study 1, while in Study 2 the fact that 11-12 year-olds had more previous experience might have led to faster learning in the current thesis. Ideally, naïve language learners of various ages would need to participate in an extended training paradigm to investigate the role of age of acquisition without conflating any previous exposure to the language. However, in reality, this might lead to recruitment difficulties as the incentive for schools to have children take time out of their class schedule to learn a completely novel language might not be accepted as readily. This makes it harder for such research to actually be performed in future.

Finally, this thesis raises a lot of questions regarding the properties of the talkers and stimuli used in the training paradigm. It is currently still unclear what information people actually pick up on when hearing a voice and applying this information in the phonetic training paradigm, and whether there are specific acoustic properties they use over others in the many features that vary across multi-talker input. Moreover, if there are particular features that learners pick up

on, does that mean that some voices are better suited for learning to generalise across talkers than others? A final question that could prove very fruitful in further research, is how many voices constitute the ideal number to provide learners with sufficient variability in the input for it to be beneficial for generalisation, while at the same time providing just enough for it to not be detrimental to the task demands. This may differ for leaners of different ages. All of these are as of yet open questions that could be pursued in further research through means of careful stimuli and paradigm design. Ultimately, the goal has to be to develop training materials that are tailored to learners of different ages, so that each age group can reach learning and generalisation to the best of their abilities.

Chapter 6. References

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

Aliaga-García, C. (2009). Effects of audiovisual auditory and articulatory training on second-language (L2) vowel category learning. *The Journal of the Acoustical Society of America*, *125*(4), 2765–2765. https://doi.org/10.1121/1.4784692

Aliaga-García, C., Mora, J. C., & Cerviño-Povedano, E. (2011). L2 speech learning in adulthood and phonological short-term memory. *Poznań Studies in Contemporary Linguistics*, *47*(1), 1. https://doi.org/10.2478/psicl-2011-0002

Alispahic, S., Escudero, P., & Mulak, K. E. (2014). Is more always better? The perception of Dutch vowels by English versus Spanish listeners. In J. Hay & E. Parnell (Eds.), *Proceedings of the 15th Australasian International Conference on Speech Science and Technology* (pp. 219–222). Retrieved from http://www.nzilbb.canterbury.ac.nz/SST.shtml

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

Alshangiti, W. (2015). *Speech production and perception in adult Arabic learners of English : A comparative study of the role of production and perception training in the acquisition of British English vowels* (PhD Thesis, University College London). Retrieved from http://discovery.ucl.ac.uk/id/eprint/1466643

Alshangiti, W., & Evans, B. G. (2014). 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. *Proceedings of the 10th International Seminar on Speech Production*. Retrieved from http://www.issp2014.uni-koeln.de/wp-content/uploads/2014/Proceedings_ISSP_revised.pdf

Antoniou, M., & Wong, P. C. M. (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

Antoniou, M., Wong, P. C. M., & Wang, S. (2015). The Effect of Intensified Language Exposure on Accommodating Talker Variability. *Journal of Speech, Language, and Hearing Research*, *58*(3), 722–727. https://doi.org/10.1044/2015_JSLHR-S-14-0259

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

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

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

Baker, R. J., & Rosen, S. (2001). Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking. *British Journal of Audiology*, *35*(1), 43–52. https://doi.org/10.1080/03005364.2001.11742730

Barcroft, J., & Sommers, M. S. (2005). Effects of Acoustic Variability on Second Language Vocabulary Learning. *Studies in Second Language Acquisition*, *27*(03), 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*(03), 423–449. https://doi.org/10.1017/S0272263113000582

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

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

Belemans, R., & Goossens, J. (2000). *Woordenboek van Brabantse Dialecten. Deel III. Inleiding en Klankgeografie*. Assen: Van Gorcum.

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* (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., McRoberts, G. W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, *109*(2), 775–794. https://doi.org/10.1121/1.1332378

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn & M. J. Munro (Eds.), *Language Experience in Second Language Speech Learning: In honor of James Emil Flege* (pp. 13–34). https://doi.org/10.1075/lllt.17.07bes

Bialystok, E., & Miller, B. (1999). The problem of age in second-language acquisition: Influences from language, structure, and task. *Bilingualism:*

Language and Cognition, *2*(2), 127–145. https://doi.org/10.1017/S1366728999000231

Birdsong, D. (Ed.). (1999). *Second Language Acquisition and the Critical Period Hypothesis*. London: Lawrence Erlbaum Associates.

Birdsong, D., & Molis, M. (2001). On the Evidence for Maturational Constraints in Second-Language Acquisition. *Journal of Memory and Language*, *44*(2), 235–249. https://doi.org/10.1006/jmla.2000.2750

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

Bolker, B. (2019). *GLMM FAQ*. Retrieved August 7, 2019, from http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#should-i-treat-factor-xxx-as-fixed-or-random

Booij, G. (1999). *The Phonology of Dutch* (J. Durand, Ed.). Oxford: Oxford University Press.

Boxtel, S. van, Bongaerts, T., & Coppen, P.-A. (2005). Native-like attainment of dummy subjects in Dutch and the role of the L1. *IRAL - International Review of Applied Linguistics in Language Teaching*, *43*(4), 355–380. https://doi.org/10.1515/iral.2005.43.4.355

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

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (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

Braginsky, M. (2018). *ggpirate: pirate plots for ggplot2* [R package]. Retrieved from https://github.com/mikabr/ggpirate

Broersma, M., & Cutler, A. (2011). Competition dynamics of second-language listening. *The Quarterly Journal of Experimental Psychology*, *64*(1), 74–95. https://doi.org/10.1080/17470218.2010.499174

Brosseau-Lapré, F., Rvachew, S., Clayards, M., & Dickson, D. (2013). Stimulus variability and perceptual learning of nonnative vowel categories. *Applied Psycholinguistics*, *34*(3), 419–441. https://doi.org/10.1017/S0142716411000750

Bürki, A., Welby, P., Clément, M., & Spinelli, E. (2019). Orthography and second language word learning: Moving beyond "friend or foe?" *The Journal of the Acoustical Society of America*, *145*(4), EL265–EL271. https://doi.org/10.1121/1.5094923

Carlet, A. (2017). *L2 perception and production of English consonants and vowels by Catalan speakers: The effects of attention and training task in a cross-training study* (PhD Thesis, Universitat Autònoma de Barcelona). Retrieved from https://core.ac.uk/download/pdf/132091676.pdf

Carlet, A., & Cebrian, J. (2015). Identification vs. discrimination training: Learning effects for trained and untrained sounds. *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow.

Carley, P., Mees, I. M., & Collins, B. (2018). *English Phonetics and Pronunciation Practice*. London: Routledge.

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*(3), 386–404. https://doi.org/10.1016/0022-0965(82)90054-6

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

Chirrey, D. (1999). Edinburgh: descriptive material. In P. Foulkes & G. Docherty (Eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 223–229). London: Arnold.

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

Collins, B., den Hollander, S. P., Mees, I. M., & Rodd, J. (2011). *Sounding Better: A Practical Guide to English Pronunciation for Speakers of Dutch*. Holten, the Netherlands: Walvaboek.

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

Collins, B., & Mees, I. M. (2012). *Accepted American Pronunciation: A Practical Guide for Speakers of Dutch*. Holten, the Netherlands: Walvaboek.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, *106*(2), 633–664. https://doi.org/10.1016/j.cognition.2007.03.013

Creel, S. C., & Bregman, M. R. (2011). How Talker Identity Relates to Language Processing. *Language and Linguistics Compass*, *5*(5), 190–204. https://doi.org/10.1111/j.1749-818X.2011.00276.x

Cruttenden, A. (2008). *Gimson's Pronunciation of English* (Seventh ed). London, UK: Hodder Education.

Cumming, G. (2014). The New Statistics. *Psychological Science*, *25*(1), 7–29. https://doi.org/10.1177/0956797613504966

Dąbrowska, E. (2019). Experience, Aptitude, and Individual Differences in Linguistic Attainment: A Comparison of Native and Nonnative Speakers. *Language Learning*, *69*(March), 72–100. https://doi.org/10.1111/lang.12323

de Bont, A. P. (1962). *Dialekt van Kempenland: Meer in het bijzonder D'Oerse Taol. Deel I: Klank- en Vormleer en Enige Syntaktische Bijzonderheden*. Assen: Van Gorcum.

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology*, *91*(3), 450–476. https://doi.org/10.1037/0022-0663.91.3.450

DeKeyser, R. M. (2000). The Robustness of Critical Period Effects in Second Language Acquisition. *Studies in Second Language Acquisition*, *22*(4), 499–533.

DeKeyser, R. M., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*(3), 413–438. https://doi.org/10.1017/S0142716410000056

Díaz, B., Mitterer, H., Broersma, M., Escera, C., & Sebastián-Gallés, N. (2016). Variability in L2 phonemic learning originates from speech-specific capabilities: An MMN study on late bilinguals. *Bilingualism: Language and Cognition*, *19*(5), 955–970. https://doi.org/10.1017/S1366728915000450

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). https://doi.org/10.1093/acprof:oso/9780199688890.003.0012

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. (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

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238. https://doi.org/10.3758/BF03206487

EP Nuffic. (n.d.). *Tweetalig primair onderwijs (tpo)*. Retrieved November 21, 2018, from https://www.nuffic.nl/onderwerpen/tweetalig-primair-onderwijs-tpo/

EP Nuffic. (2015). *TPO Brochure - English*. Retrieved November 21, 2018, from https://www.nuffic.nl/onderwerpen/tweetalig-primair-onderwijs-tpo/

Escudero, P. (2015). Orthography plays a limited role when learning the phonological forms of new words: The case of Spanish and English learners of novel Dutch words. *Applied Psycholinguistics*, *36*(1), 7–22. https://doi.org/10.1017/S014271641400040X

Escudero, P., & Boersma, P. (2002). The subset problem in L2 perceptual

development: Multiple-category assimilation by Dutch learners of Spanish. In B. Skarabela, S. Fish, & A. H.-J. Do (Eds.), *Proceedings of the 26th annual Boston University conference on language development* (pp. 208–219). Somerville: Cascadilla Press.

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. (2019a). *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

Eurostat. (2019b). *Foreign language skills 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.

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

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. (2002). Interactions between native and second-language phonetic

systems. *An Integrated View of Language Development: Papers in Honor of Henning Wode*, 217–244.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*(6), 975–991. https://doi.org/10.3758/s13423-012-0322-y

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

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

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement* [R package]. Retrieved from https://cran.r-project.org/package=irr

Ganugapati, D., & Theodore, R. M. (2018). Structured phonetic variation facilitates talker identification. *The Journal of the Acoustical Society of America*, *144*(3), 1798–1798. https://doi.org/10.1121/1.5067932

Gathercole, S. E. (1999). Cognitive approaches to the development of short-term memory. *Trends in Cognitive Sciences*, *3*(11), 410–419. https://doi.org/10.1016/S1364-6613(99)01388-1

Gathercole, S. E., & Baddeley, A. D. (1993). Phonological working memory: A critical building block for reading development and vocabulary acquisition? *European Journal of Psychology of Education*, *8*(3), 259–272.

https://doi.org/10.1007/BF03174081

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*(2), 103–127. https://doi.org/10.1080/09658219408258940

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

Gong, J., Yang, Z., Ji, X., & Wang, F. (2019). Investigating the effectiveness of auditory training on Chinese listeners' perception of English vowels. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Goossens, J. (1981). Middelnederlandse vocaalsystemen. *Bijlagen van de Vereniging Voor Limburgse Dialect- En Naamkunde*, *2* ([Overdruk uit 'Verslagen en Mededelingen' van de Koninklijke Academie voor Nederlandse Taal-en Letterkunde, jg. 1980, afl. 2]). Retrieved from https://dbnl.org/tekst/goos003midd01_01/

Goriot, C. (2019). *Early-English education works no miracles Cognitive and*

linguistic development of mainstream, early-English, and bilingual primary-school pupils in the Netherlands. (PhD Thesis, Radboud University Nijmegen). Retrieved from https://www.lotpublications.nl/early-english-education-works-no-miracles

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

Grünwald, P. D., & de Heide, R. (2018). Invited discussion to the paper Using Stacking to Average Bayesian Predictive Distributions. *Bayesian Analysis*, *13*(3), 917–1007. https://doi.org/10.1214/17-ba1091

Gussenhoven, C. (2002). Dutch. In *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet* (pp. 74–77). Cambridge: Cambridge University Press (CUP).

Gussenhoven, C. (2007). Wat is de beste transcriptie voor het Nederlands? *Nederlandse Taalkunde*, *12*, 331–350. Retrieved from http://gep.ruhosting.nl/carlos/wat_is_de_beste_transcriptie.pdf

Gussenhoven, C., & Broeders, A. (1997). *English pronunciation for student teachers*. Groningen: Wolters-Noordhoff.

Harrington, J. (2006). An acoustic analysis of 'happy-tensing' in the Queen's Christmas broadcasts. *Journal of Phonetics*, *34*(4), 439–457. https://doi.org/10.1016/j.wocn.2005.08.001

Harrington, J., Palethorpe, S., & Watson, C. I. (2000). Does the Queen speak the Queen's English? *Nature*, *408*(6815), 927–928. https://doi.org/10.1038/35050160

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*(April), 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Hattori, K., & Iverson, P. (2009). English /r/-/l/ category assimilation by Japanese adults: Individual differences and the link to identification accuracy. *The Journal of the Acoustical Society of America*, *125*(1), 469–479. https://doi.org/10.1121/1.3021295

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

Hayes-Harb, R., Nicol, J., & Barker, J. (2010). Learning the Phonological Forms of New Words: Effects of Orthographic and Auditory Input. *Language and Speech*, *53*(3), 367–381. https://doi.org/10.1177/0023830910371460

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, *28*(4), 377–396. https://doi.org/10.1006/jpho.2000.0121

Hazan, V., Messaoud-Galusi, S., Rosen, S., Nouwens, S., & Shakespeare, B. (2009). Speech Perception Abilities of Adults With Dyslexia: Is There Any Evidence for a True Deficit? *Journal of Speech, Language, and Hearing Research*, *52*(6), 1510–1529. https://doi.org/10.1044/1092-4388(2009/08-0220)

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication*, *47*(3), 360–378. https://doi.org/10.1016/j.specom.2005.04.007

Heeren, W. F. L., & Schouten, M. E. H. (2008). Perceptual development of phoneme contrasts: How sensitivity changes along acoustic dimensions that contrast phoneme categories. *The Journal of the Acoustical Society of America*, *124*(4), 2291–2302. https://doi.org/10.1121/1.2967472

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

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

Holliday, J. J. (2016). Second Language Experience Can Hinder the Discrimination of Nonnative Phonological Contrasts. *Phonetica*, *73*(1), 33–51. https://doi.org/10.1159/000443312

Huckvale, M. (2016). *ProRec: Prompt & Record* [Software]. Retrieved from http://www.phon.ucl.ac.uk/resource/prorec/

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

Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *Journal of Phonetics*, *52*, 105–120. https://doi.org/10.1016/j.wocn.2015.06.007

Hwang, H., & Lee, H.-Y. (2015). The effect of high variability phonetic training on the production of English vowels and consonants. *Proceedings of the 18th International Congress of Phonetic Sciences*. Retrieved from https://pdfs.semanticscholar.org/b1de/abb7eb1385c1731895f1a4a7fc5a9d144ce3.pdf

Ingvalson, E. M., Barr, A. M., & Wong, P. C. M. (2013). Poorer Phonetic Perceivers Show Greater Benefit in Phonetic-Phonological Speech Learning. *Journal of Speech Language and Hearing Research*, *56*(3), 1045. https://doi.org/10.1044/1092-4388(2012/12-0024)

Ioup, G. (2008). 2. Exploring the role of age in the acquisition of a second language

phonology. In J. G. Hansen Edwards & M. L. Zampini (Eds.), *Phonology and Second Language Acquisition* (pp. 41–62). https://doi.org/10.1075/sibil.36.04iou

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*(5), 1305–1316. https://doi.org/10.1037/0096-1523.34.5.1305

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

Iverson, P., Kuhl, P. K., Akahane-yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2001). A perceptual interference account of acquisition difficulties for non-native phonemes. *Speech, Hearing and Language: Work in Progress*, *13*, 106–118.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57.

https://doi.org/10.1016/S0010-0277(02)00198-1

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145–160. https://doi.org/10.1017/S0142716411000300

Iverson, P., & Preece-Pinet, M. (2008). Training English vowels for French speakers with varying English experience. *The Journal of the Acoustical Society of America*, *123*(5), 3734. https://doi.org/10.1121/1.2935238

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

Jamieson, D. G., & Morosan, D. E. (1989). Training new, nonnative speech contrasts: a comparison of the prototype and perceptual fading techniques. *Canadian Journal of Psychology*, *43*(1), 88–96. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2819599

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

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99. https://doi.org/10.1016/0010-0285(89)90003-0

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds. *The Journal of the*

Acoustical Society of America, *138*(2), 817–832. https://doi.org/10.1121/1.4926561

Kartushina, N., & Martin, C. D. (2019). Talker and Acoustic Variability in Learning to Produce Nonnative Sounds: Evidence from Articulatory Training. *Language Learning*, *69*(1), 71–105. https://doi.org/10.1111/lang.12315

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. https://doi.org/10.3758/BRM.42.3.627

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kim, Y. H., & Hazan, V. (2010). Individual variability in the perceptual learning of L2 speech sounds and its cognitive correlates. *New Sounds 2010. Sixth International Symposium on the Acquisition of Second Language Speech*, 251–256. Retrieved from http://ifa.amu.edu.pl/newsounds/files/proceedings/proceedings_quotable_version.pdf

Kirstein, A. (2018). *Schwa epenthesis and schwa deletion.* Retrieved December 5, 2017, from Taalportaal website: http://www.taalportaal.org/taalportaal/topic/pid/topic_20150914151659461

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857. https://doi.org/10.1121/1.398894

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), 148–203. https://doi.org/10.1037/a0038695

Kloots, H., De Schutter, G., Gillis, S., & Swerts, M. (2002). Sjwa-insertie in eindclusters: variatiepatronen in het Standaardnederlands. *Nederlandse Taalkunde*, *7*(2), 97–124.

Kloots, H., Gillis, S., De Maeyer, S., & Verhoeven, J. (2012). De duur van de svarabhaktivocaal in het Standaardnederlands: Een pioniersstudie. *Tijdschrift Voor Nederlandse Taal- En Letterkunde*, (January), 1–18. Retrieved from http://openaccess.city.ac.uk/7458/

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81. https://doi.org/10.1016/j.cognition.2007.07.013

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268. https://doi.org/10.3758/BF03193841

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*(1), 1–15. https://doi.org/10.1016/j.jml.2006.07.010

Kuhl, P. K. (1993). Innate Predispositions and the Effects of Experience in Speech Perception: The Native Language Magnet Theory. In *Developmental Neurocognition: Speech and Face Processing in the First Year of Life* (pp. 259–274). https://doi.org/10.1007/978-94-015-8234-6_22

Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., &

Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 979–1000. https://doi.org/10.1098/rstb.2007.2154

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

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. https://doi.org/10.2307/2529310

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

Lengeris, A. (2009). Perceptual Assimilation and L2 Learning: Evidence from the Perception of Southern British English Vowels by Native Speakers of Greek and Japanese. *Phonetica*, *66*(3), 169–187. https://doi.org/10.1159/000235659

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*(6), 3757–3768. https://doi.org/10.1121/1.3506351

Lenneberg, E. H. (1967). *Biological Foundations of Language*. Wiley.

Lev-Ari, S. (2017). Talking to fewer people leads to having more malleable linguistic representations. *PLOS ONE*, *12*(8), e0183593.

https://doi.org/10.1371/journal.pone.0183593

Levitt, H. (1971). Transformed Up-Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

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

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of the Acoustical Society of America*, *96*(4), 2076–2087. https://doi.org/10.1121/1.410149

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

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

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

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157. https://doi.org/10.1016/0010-0285(88)90017-5

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. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2526857

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

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

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

McCafferty, K. (1999). (London)Derry: between Ulster and local speech - class, ethnicity and language change. In P. Foulkes & G. Docherty (Eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 246–264). London: Arnold.

McCarthy, K. M., Mahon, M., Rosen, S., & Evans, B. G. (2014). Speech Perception and Production by Sequential Bilingual Children: A Longitudinal Study of Voice Onset Time Acquisition. *Child Development*, *85*(5). https://doi.org/10.1111/cdev.12275

McMurray, B., Aslin, R. N., Tanenhaus, M. K., Spivey, M. J., & Subik, D. (2008). Gradient sensitivity to within-category variation in words and syllables. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1609–1631. https://doi.org/10.1037/a0011747

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. https://doi.org/10.1016/S0010-0277(02)00157-9

Mees, I., & Collins, B. (1983). A phonetic description of the vowel system of Standard Dutch (ABN). *Journal of the International Phonetic Association*, *13*(2), 64–75. https://doi.org/10.1017/S0025100300002565

Mitterer, H., & Ernestus, M. (2008). The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition*, *109*(1), 168–173. https://doi.org/10.1016/j.cognition.2008.08.002

Morosan, D. E., & Jamieson, D. G. (1989). Evaluation of a technique for training new speech contrasts: generalization across voices, but not word-position or task. *Journal of Speech and Hearing Research*, *32*(3), 501–511. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2779195

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

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, *38*(1), 1–21. https://doi.org/10.1111/1469-8986.3810001

Nimz, K., & Khattab, G. (2019). On the role of orthography in L2 vowel production: The case of Polish learners of German. *Second Language Research*, 026765831982842. https://doi.org/10.1177/0267658319828424

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

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)

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration Becoming the Norm in Psychological Science. *APS Observer*, *31*(3). Retrieved from https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech Perception as a Talker-Contingent Process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Palfi, B., & Dienes, Z. (2019). Why Bayesian "evidence for H1" in one condition and "evidence for H0" in another does not mean Bayesian evidence for a difference between conditions: The role of Bayes factors in testing interactions. *PREPRINT*. https://doi.org/10.31234/osf.io/qjrg4

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

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. Department of Applied Mathematics and Theoretical Physics, Cambridge England, *Technical report NA2009/06*.

R Core Team. (2018). *R: A Language and Environment for Statistical Computing* [Software]. Retrieved from https://www.r-project.org

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

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: insights from a multiple case study of dyslexic adults. *Brain*, *126*(4), 841–865. https://doi.org/10.1093/brain/awg076

Rato, A. (2014). Effects of perceptual training on the perception (and production) of English vowels by native speakers of European Portuguese. *Concordia Working Papers in Applied Linguistics*, *5*, 529–546. Retrieved from http://doe.concordia.ca/copal/volumes/

Rato, A., & Rauber, A. (2015). The Effects of Perceptual Training on the Production of English vowel contrasts by Portuguese Learners. *Proceedings of the 18th International Congress of Phonetic Sciences*. Retrieved from http://www.icphs2015.info/pdfs/Papers/ICPHS0656.pdf

Ricketts, J., Bishop, D. V. M., & Nation, K. (2009). Orthographic Facilitation in Oral Vocabulary Acquisition. *Quarterly Journal of Experimental Psychology*, *62*(10), 1948–1966. https://doi.org/10.1080/17470210802696104

Roach, P. (1991). *English Phonetics and Phonology: A Practical Course* (Second edi). Cambridge: Cambridge University Press.

Rodd, J. M. (2018). *Ensuring data quality when you can't see your participants*. [Video]. London, UK. Retrieved from https://www.youtube.com/watch?v=wJZ3MRkFb_I

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological

processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the Signal by Adding Noise: The Role of Noncontrastive Phonetic Variability in Early Word Learning. *Infancy*, *15*(6), 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Rouder, J. N. (2019). On The Interpretation of Bayes Factors: A Reply to de Heide and Grunwald. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/m6dhw

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

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

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

Scheltinga, F. (1998). *Nonwoord Repetitie Test*. Rotterdam: CED-Groep.

Scobbie, J. M., Gibbon, F., Hardcastle, W. J., & Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In M. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 194–207). Retrieved from https://eresearch.qmu.ac.uk/handle/20.500.12289/3352

Scott Sheldon, P. (1974). *Spoken English*. Antwerp, Belgium: De Nederlandsche Boekhandel.

Sebregts, K. D. C. J. (2015). *The sociophonetics and phonology of Dutch 'r.'* (PhD Thesis, Universiteit Utrecht). Retrieved from https://www.lotpublications.nl/the-sociophonetics-and-phonology-of-dutch-r

Shaw, J. A., Best, C. T., Docherty, G., Evans, B. G., Foulkes, P., Hay, J., & Mulak, K. E. (2018). Resilience of English vowel perception across regional accent variation. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *9*(1), 11. https://doi.org/10.5334/labphon.87

Shinohara, Y. (2014). *Perceptual Training of English /r/ and /l/ for Japanese Adults, Adolescents and Children*. (PhD Thesis, University College London). Retrieved from https://discovery.ucl.ac.uk/id/eprint/1421176

Shinohara, Y., & Iverson, P. (2013). Computer-based English /r/-/l/ perceptual training for Japanese children. *Proceedings of Meetings on Acoustics*, *19*, 060049–060049. https://doi.org/10.1121/1.4800136

Shinohara, Y., & Iverson, P. (2015). Effects of English /r/-/l/ Perceptual Training on Japanese Children's Production. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: the University of Glasgow. ISBN 978-0-85261-941-4. Paper number 1041.1-9 retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0540.pdf

Shinohara, Y., & Iverson, P. (2018). High variability identification and discrimination training for Japanese speakers learning English /r/–/l/. *Journal of Phonetics*, *66*, 242–251. https://doi.org/10.1016/j.wocn.2017.11.002

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

Simon, E., Chambless, D., & Kickhöfel Alves, U. (2010). Understanding the role of orthography in the acquisition of a non-native vowel contrast. *Language Sciences*, *32*(3), 380–394. https://doi.org/10.1016/j.langsci.2009.07.001

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

Snowling, M., & Hulme, C. (1994). The development of phonological skills. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *346*(1315), 21–27. https://doi.org/10.1098/rstb.1994.0124

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

Sommers, M. S., & Barcroft, J. (2007). An integrated account of the effects of acoustic variability in first language and second language: Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics*, *28*(2), 231–249. https://doi.org/10.1017/S0142716407070129

Sommers, M. S., & Barcroft, J. (2011). Indexical information, encoding difficulty, and second language vocabulary learning. *Applied Psycholinguistics*, *32*(02), 417–434. https://doi.org/10.1017/S0142716410000469

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. https://doi.org/10.1038/41102

Strange, W. (1995). *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (W. Strange, Ed.). Timonium, MD: York Press.

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

Strycharczuk, P., & Scobbie, J. M. (2017). Fronting of Southern British English high-back vowels in articulation and acoustics. *The Journal of the Acoustical Society of America*, *142*(1), 322–331. https://doi.org/10.1121/1.4991010

Stuart-Smith, J. (1999). Glasgow: accent and voice quality. In P. Foulkes & G. Docherty (Eds.), *Urban Voices: Accent Studies in the British Isles* (pp. 203–222). London: Arnold.

Swanenberg, J. (2009). *Van alterande sorte: Brabants tussen dialect en standaardtaal* [Inaugural lecture]. Retrieved from https://www.erfgoedbrabant.nl/media/220836/inaugurele-rede-jos-swanenberg.pdf

Swanenberg, J., & Brok, H. (2008). *Het Brabants beschreven*. Veerhuis.

Swanenberg, J., & Swanenberg, C. (2002). *Oost-Brabants*. Den Haag: Sdu Uitgevers.

Taimi, L., Jähi, K., Alku, P., & Peltola, M. S. (2014). Children Learning a Non-native Vowel – The Effect of a Two-day Production Training. *Journal of Language Teaching and Research*, *5*(6), 1229–1235. https://doi.org/10.4304/jltr.5.6.1229-1235

Thomson, R. I. (2011). Computer Assisted Pronunciation Training: Targeting Second Language Vowel Perception Improves Pronunciation. *CALICO Journal*, *28*(3), 744–765. https://doi.org/10.11139/cj.28.3.744-765

TNS Opinion & Social. (2012). *Special Eurobarometre 386: Europeans and their languages*. Retrieved from

http://ec.europa.eu/commfrontoffice/publicopinion/archives/eb_special_399_380_en.htm

Treiman, R., & Kessler, B. (2014). *How Children Learn to Write Words*. Oxford, UK: Oxford University Press.

van de Velde, H., & van Hout, R. (1999). The Pronunciation of (r) in Standard Dutch. *Linguistics in the Netherlands*, *16*, 177–188. https://doi.org/10.1075/avt.16.16van

van de Velde, H., & van Hout, R. (Eds.). (2001). *'r-atics: Sociolinguistic, phonetic and phonological characteristics of /r/*.

van der Haagen, M. (1998). *Caught between norms: The English pronunciation of Dutch learners*. (PhD Thesis, Katholieke Universiteit Nijmegen). Holland Academic Graphics, The Hague.

Vanhove, J. (2013). The Critical Period Hypothesis in Second Language Acquisition: A Statistical Critique and a Reanalysis. *PLoS ONE*, *8*(7), e69172. https://doi.org/10.1371/journal.pone.0069172

Verstraeten, B., & van de Velde, H. (2001). Socio-geographical variation of /r/ in standard Dutch. In H. van de Velde & R. van Hout (Eds.), *'r-atics: Sociolinguistic, phonetic and phonological characteristcs of /r/* (pp. 45–62). Université Libre de Bruxelles.

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of Acoustic Variability in the Perceptual Learning of Non-Native-Accented Speech Sounds. *Phonetica*, *64*(2–3), 122–144. https://doi.org/10.1159/000107913

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

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. *ICPhS Conference Proceedings*, 1537–1540. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2003/p15_1537.html

Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. (1999). Training American listeners to perceive Mandarin tones. *The Journal of the Acoustical Society of America*, *106*(6), 3649–3658. https://doi.org/10.1121/1.428217

Warner, N. (1999). Timing of perception of vocalic distinctive features: implications for vowel system universals. *XIVth International Congress of Phonetic Sciences (ICPhS99)*, 1961–1964. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS1999/p14_1055.html

Warner, N., Jongman, A., Cutler, A., & Mücke, D. (2001). The phonological status of Dutch epenthetic schwa. *Phonology*, *18*(3), 387–420. https://doi.org/10.1017/S0952675701004213

Weatherholtz, K. (2015). *Perceptual Learning of Systemic Cross-Category Vowel Variation*. (PhD thesis, Ohio State University). Retrieved from http://rave.ohiolink.edu/etdc/view?acc_num=osu1429782580

Wells, J. C. (1982a). *Accents of English 1: An Introduction*. Cambridge: Cambridge University Press.

Wells, J. C. (1982b). *Accents of English 2: The British Isles*. Cambridge: Cambridge University Press.

Wells, J. C. (1982c). *Accents of English 3: Beyond the British Isles*. Cambridge: Cambridge University Press.

Werker, J. F. (1995). Age-Related Changes in Cross-Language Speech Perception:

Standing at the Crossroads. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 155–169). Baltimore: York Press.

Werker, J. F., & Curtin, S. (2005). PRIMIR: A Developmental Framework of Infant Speech Processing. *Language Learning and Development*, *1*(2), 197–234. https://doi.org/10.1207/s15473341lld0102_4

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1). https://doi.org/10.18637/jss.v040.i01

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

Willems, N. (1982). *English Intonation from a Dutch Point of View* (2010 repri). Dordrecht: De Gruyter Mouton.

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

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

Wong, J. W. S. (2015a). Comparing the effects of perception and production training on the learning of English vowel contrast /e/ and /æ/ by Cantonese ESL learners. *Proceedings of the 18th International Congress of Phonetic Sciences*. Retrieved from

https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0943.pdf

Wong, J. W. S. (2015b). The Effects High-Variability Phonetic Training on Cantonese ESL Learners' Production of English Vowel Contrasts - An Acoustic Analysis. *Phonetics Teaching and Learning Conference*, 107–111. Retrieved from https://repository.hkbu.edu.hk/hkbu_staff_publication/6235/

Wong, P. C. M., & Perrachione, T. K. (2007). Learning pitch patterns in lexical identification by native English-speaking adults. *Applied Psycholinguistics*, *28*(4), 565–585. https://doi.org/10.1017/S0142716407070312

Xie, Y. (2018). *knitr: A General-Purpose Package for Dynamic Report Generation in R* [R package]. Retrieved from https://cran.r-project.org/package=knitr

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

Ylinen, S., Uther, M., Latvala, A., Vepsäläinen, S., Iverson, P., Akahane-Yamada, R., & Näätänen, R. (2010). Training the Brain to Weight Speech Cues Differently: A Study of Finnish Second-language Users of English. *Journal of Cognitive Neuroscience*, *22*(6), 1319–1332. https://doi.org/10.1162/jocn.2009.21272

Zhu, H. (2018). *kableExtra: Construct Complex Table with "kable" and Pipe Syntax*. Retrieved from https://cran.r-project.org/package=kableExtra

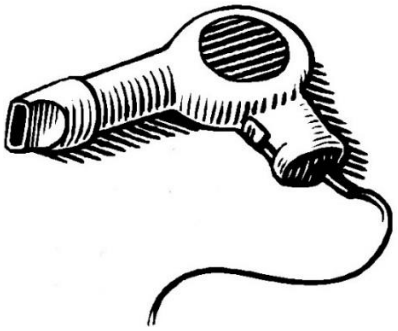
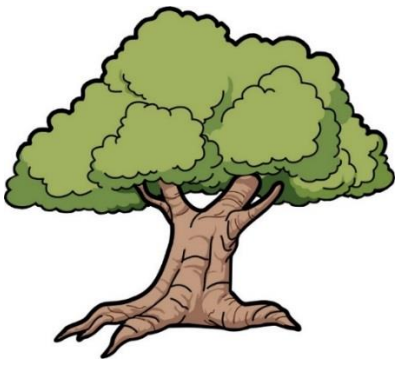
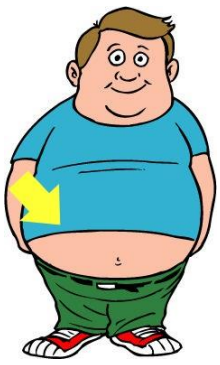
Appendix I. GOOSE-fronting formant measures

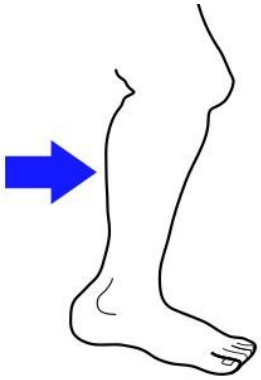
See the table below for the formant measures taken from the 4 different speakers used for the training stimuli of Study 1. Averages before coda-/l/, where GOOSE-fronting is unlikely to happen (Strycharczuk & Scobbie, 2017), are much closer to the reference values (Deterding, as cited in Cruttenden (2008)). The pre-coronal contexts show a much higher F2, often midway between values expected for /u/ and /i/, indicating significant fronting.

	Female 4		Female 5		Male 1		Male 2	
	F1	F2	F1	F2	F1	F2	F1	F2
fool	331	834	453	693	567	2143	454	1915
	325	810	443	797	732	2415	355	1430
pool	330	832	430	769	382	781	299	741
	318	828	374	772	592	1077	320	801
Luke	343	1976	383	1955	391	1606	228	1765
	328	2001	384	2016	1374	2175	293	1601
suit	355	2127	342	2303	332	1765	289	2046
	321	2058	360	2212	1658	1907	295	2004
AVERAGE pre-coronal	337	2040	367	2122	939	1863	276	1854
AVERAGE pre coda-/l/	326	826	425	758	568	1604	357	1222
REFERENCE /u/	339	1396	339	1396	302	1131	302	1131
REFERENCE /i/	319	2723	319	2723	275	2221	275	2221

Appendix II. Study 1 - Experiment 2 - Stimuli pictures

Experiment 2 and 3 real-word minimal pair items. Items are shown with the Dutch word in capital letters, and the English meaning in quotation marks after.

1	FAUN - 'faun' 	FOHN - 'hair dryer' 
2	KOUS - 'long sock' 	KEUS - 'choice' 
3	BEUK -'birch tree' 	BUIK –'stomach' 

4	TEUN - ‘John’ 	TUIN -‘garden’ 
5	KOUD - ‘cold’ 	KUIT - ‘calf’ 
6	ZOUT - ‘salt’ 	ZUID - ‘south’ 

Appendix III. Bayes Factor computation and justification

As noted in the text, an estimation of the effect from independent data was used wherever possible. However, in some cases this was not possible and it was necessary to base the estimate using a value from within the same data set. For this, a set of justifications were used. Each individual model is laid out below, specifying the calculation of the H1 (the predicted effect x) and what justification was used. The labelled justifications refer to the Justifications described below those tables.

Study 1

EXPERIMENT 1

ADULTS

Exp1_Adult_DM

EFFECT	H1	COMPUTATION	JUSTIFICATION
Comparison with Chance	2.098	Exp2_LabAdults_DM Intercept - chance	Equivalent effect from independent participants
Session	1.539	(logodds(.99) - logodds(mean(.82,.70, .88,.88))) /2 = 1.539386	A measure of maximal possible increase from pre- to post-test, based on difference between average performance at pre-test for this age group (across the experiments) subtracted from ceiling value (99%).
Novelty*session	1.539		
Variability*session	1.539		
Variability*novelty *session	1.539		

Exp1_Adult_Train

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.102	Exp1_child_train intercept - chance	Equivalent effect from independent participants
Block	0.139	Exp1_adult_train intercept - chance) / 3 = 0.418/3	Justification B e = Exp1_adult_train intercept - chance a = block, 4 levels
Variability*block	0.063	Exp1_adult_train block	Justification A a = condition, a1 = LV, a2 = HV e = Exp1_adult_train block

Exp1_Adult_ID

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.418	Exp1_adult_Train intercept - chance	In the absence of an equivalent effect in children, and since there is no data from a 3AFC ID test in any of the other experiments, base the estimate on the value from the same participants in the training task (which is also 3AFC).
Variability	0.418	Exp1_Adult_ID intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept
Novelty	0.418	Exp1_Adult_ID intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept
Variability*novelty	0.317	(Exp1_adult_ID intercept - chance) * (2/3) = 0.475 *2/3	*Justification C* a = novelty, a1 = untrained, a2 = trained b = condition, b1 = LV, b2 = HV e = Exp1_adult_Train equivalent- chance (note, we had planned to base it on effect of overall variability in the model (Justification A) but this effect isn't seen in the current data)

CHILDREN

Exp1_Child_DM

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	1.983	Exp2_Child_DM intercept - chance	Equivalent effect from independent participants
Session	1.665	(logodds(.99) - logodds(mean(.78,.73,.74)))/2= 1.664727	A measure of maximal possible increase from pre- to post-test, based on difference between average performance at pre-test for this age group (across the experiments) subtracted from ceiling value (99%).
Novelty*session	1.665		
Variability*session	1.665		
Variability*novelty*session	1.665		

Exp1_Child_Train

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.418	Exp1_adult_train intercept - chance	Equivalent effect from independent participants
Block	0.189	Exp1_adult_train block	Equivalent effect from independent participants
Variability*block	0.189	Exp1_adult_train block	*Justification A* a = condition, a1 = LV, a2 = HV e = Exp1_adult_train block (note, we had planned to use effect of block in the current model but this effect isn't seen in the current data)

Exp1_Child_ID

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.475	Exp1_adult_ID intercept - chance	Equivalent effect from independent participants
Variability	0.475	Exp1_adult_ID intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = intercept (note, we can't use intercept from current data as it is negative, estimate is based on the adult intercept instead)
Novelty	0.475	Exp1_adult_ID intercept - chance	*Justification A* a = novelty, a1 = untrained, a2 = trained e = intercept (note, we can't use intercept from current data as it is negative, estimate is based on the adult intercept instead)
Variability*novelty	0.317	(Exp1_adult_ID intercept – chance) * (2/3) = 0.475 * (2/3)	*Justification C* a = novelty, a1 = untrained, a2 = trained b = condition, b1 = LV, b2 = HV e = intercept (note, we can't use intercept from current data as it is negative, estimate is based on the adult intercept instead)

AGE COMPARISON

Model	Effect	H1	COMPUTATION	JUSTIFICATION
Exp1_AgeComp_DM	Age group	1.204	Exp2_LabAdult_VS_Child age group	Equivalent effect from independent participants
Exp1_AgeComp_Train	Age group	0.257	Exp1_AgeComp_Train intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept
	Group by block	0.093	Exp1_AgeComp_Train block	*Justification A* a = condition, a1 = LV, a2 = HV e = block
Exp1_AgeComp_ID	Age group	0.226	Exp1_AgeComp_ID intercept – chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

EXPERIMENT 2

ADULTS

Exp2_LabAdult_DM

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.475	Exp1_Adult_DM intercept - chance	Equivalent effect from independent participants
Session	1.539	(logodds(.99) - logodds(mean(.82,.70,.88,.88))) /2 = 1.539386	A measure of maximal possible increase from pre- to post-test, based on difference between average performance at pre-test for this age group (across the experiments) subtracted from ceiling value (99%).
Novelty*session	1.539		
Variability*session	1.539		
Variability*novelty*session	1.539		

Exp2_OnlineAdult_DM

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.475	Exp1_Adult_DM intercept - chance	Equivalent effect from independent participants
Session	1.539	(logodds(.99) - logodds(mean(.82,.70,.88,.88))) /2 = 1.539386	A measure of maximal possible increase from pre- to post-test, based on difference between average performance at pre-test for this age group (across the experiments) subtracted from ceiling value (99%).
Novelty*session	1.539		
Variability*session	1.539		
Variability*novelty*session	1.539		

Exp2_LabAdults_Train

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.544	Exp2_OnlineAdult_Train intercept – chance	Equivalent effect from independent participants
Block	0.07	Exp2_OnlineAdult_Train block	Equivalent effect from independent participants
Variability*block	0.131	Exp2_OnlineAdult_Train variability : block	Equivalent effect from independent participants

Exp2_OnlineAdults_Train

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.795	Exp2_LabAdult_Train intercept - chance	Equivalent effect from independent participants
Block	0.211	Exp2_LabAdult_Train block	Equivalent effect from independent participants
Variability*block	0.211	Exp2_OnlineAdults_Train intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

Exp2_LabAdult_ID_minpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.584	Exp2_OnlineAdult_ID_minpair intercept - chance	Equivalent effect from independent participants
Variability	0.584	Exp2_LabAdult_ID_minpair intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

Exp2_LabAdult_ID_nonpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	3.661	Exp2_OnlineAdult_ID_nonpair intercept - chance	Equivalent effect from independent participants
Variability	0.101	Exp2_LabAdult_ID_nonpair intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

Exp2_OnlineAdult_ID_minpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.851	Exp2_LabAdult_ID_minpair intercept - chance	Equivalent effect from independent participants
Variability	0.137	Exp2_OnlineAdult_ID_minpair intercept - chance	*Justification* A a = condition, a1 = LV, a2 = HV e = adjusted intercept

Exp2_OnlineAdult_ID_nonpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	6.146	Exp2_LabAdult_ID_nonpair intercept - chance	Equivalent effect from independent participants
Variability	3.603	Exp2_OnlineAdult_ID_nonpair intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

CHILDREN

Exp2_Child_DM

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	2.029	Exp1_Child_DM intercept - chance	Equivalent effect from independent participants
Session	1.665	(logodds(.99) - logodds(mean(.78,.73,.74))) / 2 = 1.664727	A measure of maximal possible increase from pre- to post-test, based on difference between average performance at pre-test for this age group (across the experiments) subtracted from ceiling value (99%).
Novelty*session	1.665		
Variability*session	1.665		
Variability*novelty*session	1.665		

Exp2_Child_Train

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.544	Exp2_OnlineAdult_Train intercept - chance	Equivalent effect from independent participants
Block	0.07	Exp2_OnlineAdult_Train block	Equivalent effect from independent participants
Variability*block	0.131	Exp2_OnlineAdult_Train variability:block	Equivalent effect from independent participants

Exp2_Child_ID_minpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.584	Exp2_OnlineAdult_ID_minpair intercept - chance	Equivalent effect from independent participants
Variability	0.584	Exp2_Child_ID_minpair intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

Exp2_Child_ID_nonpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	3.661	Exp2_OnlineAdult_ID_nonpair equivalent	Equivalent effect from independent participants
Variability	0.101	Exp2_Child_ID_nonpair intercept	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

AGE COMPARISON

LAB ADULTS vs CHILDREN

Model	Effect	H1	COMPUTATION	JUSTIFICATION
Exp2_LabAdultvs Child_DM	Age group	0.341	Exp1_AgeComp_DM age group	Equivalent effect from independent participants
Exp2_LabAdultvs Child_Train	Age group	0.086	Exp3_AgeComp_Train age group	Equivalent effect from independent participants
	Group by block	0.186	Exp2_LabvsChild_Train intercept) / 3 = 0.557 / 3	*Justification B* e = adjusted intercept a = block, 4 levels
Exp2_LabAdultvs Child_ID_minpair	Age group	0.077	Exp3_AgeComp_ID _minp age group	Equivalent effect from independent participants
Exp2_LabAdultvs Child_ID_nonpair	Age group	0.267	Exp3_AgeComp_ID _nonp age group	Equivalent effect from independent participants

ONLINE ADULTS vs CHILDREN

Model	Effect	H1	COMPUTATION	JUSTIFICATION
Exp2_OnlineAdultvs Child_DM	Age group	0.341	Exp1_AgeComp_DM age group	Equivalent effect from independent participants
Exp2_OnlineAdultvs Child_Train	Age group	0.086	Exp3_AgeComp_Train age group	Equivalent effect from independent participants
	Group by block	0.029	Exp2_LabvsChild_Train age group: block	Equivalent effect from independent participants
Exp2_OnlineAdultvs Child_ID_minpair	Age group	0.077	Exp3_AgeComp_ID _minp age group	Equivalent effect from independent participants
Exp2_OnlineAdultvs Child_ID_nonpair	Age group	0.267	Exp3_AgeComp_ID _nonp age group	Equivalent effect from independent participants

LAB ADULTS VS ONLINE ADULTS

Model	Effect	H1	COMPUTATION	JUSTIFICATION
Exp2_LabAdultvs OnlineAdult_DM	Age group	0.341	Exp1_AgeComp_DM age group	Equivalent effect from independent participants
Exp2_LabAdultvs OnlineAdult_Train	Age group	0.086	Exp3_AgeComp_Train age group	Equivalent effect from independent participants
	Group by block	0.029	(Exp2_AgeComp_Train intercept – chance) / 3	*Justification B* e = adjusted intercept a = block, 4 levels
Exp2_LabAdultvs OnlineAdult_ID _minpair	Age group	0.077	Exp3_AgeComp_ID_minp age group	Equivalent effect from independent participants
Exp2_LabAdultvs OnlineAdult_ID _nonpair	Age group	0.267	Exp3_AgeComp_ID_nonp age group	Equivalent effect from independent participants

EXPERIMENT 3

ADULTS

Exp3_Adult_DM

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	3.316	Exp2_OnlineAdult_DM intercept - chance	Equivalent effect from independent participants
Session	1.539	(logodds(.99) - logodds(mean(.82,.70, .88,.88))) /2 = 1.539386	A measure of maximal possible increase from pre- to post-test, based on difference between average performance at pre-test for this age group (across the experiments) subtracted from ceiling value (99%).
Novelty*session	1.539		
Variability*session	1.539		
Variability*novelty*session	1.539		

Exp3_Adult_Train

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.544	Exp2_OnlineAdult_Train intercept - chance	Equivalent effect from independent participants
Block	0.07	Exp2_OnlineAdult_Train block	Equivalent effect from independent participants
Variability*block	0.131	Exp2_OnlineAdult_Train variability:block	Equivalent effect from independent participants

Exp3_Adult_ID_minpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.584	Exp2_OnlineAdult_ID_minpair intercept - chance	Equivalent effect from independent participants
Variability	0.584	Exp3_Adult_ID_minpair intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

Exp3_Adult_ID_nonpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	3.661	Exp2_OnlineAdult_ID_nonpair intercept - chance	Equivalent effect from independent participants
Variability	0.101	Exp3_Adult_ID_nonpair intercept	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

CHILDREN

Exp3_Child_DM

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	1.983	Exp2_Child_DM intercept - chance	Equivalent effect from independent participants
Session	1.665	(logodds(.99) - logodds(mean(.78,.73, .74)))/2= 1.664727	A measure of maximal possible increase from pre- to post-test, based on difference between average performance at pre-test for this age group (across the experiments) subtracted from ceiling value (99%).
Novelty*session	1.665		
Variability*session	1.665		
Variability*novelty*session	1.665		

Exp3_Child_Train

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.359	Exp2_Child_Train intercept - chance	Equivalent effect from independent participants
Block	0.083	Exp2_Child_Train block	Equivalent effect from independent participants
Variability*block	0.037	Exp2_Child_Train variability:block	Equivalent effect from independent participants
FOLLOW-UP Block:conditionLV	0.045	Exp3_Child_Train_intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept
FOLLOW-UP Block:conditionHV	0.045	Exp3_Child_Train_intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept

Exp3_Child_ID_minpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.298	Exp2_Child_ID_minpair intercept – chance	Equivalent effect from independent participants
Variability	0.298	Exp2_Child_ID_minpair intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept (note, can't use the intercept from within the current model as it is negative.)

xp3_Child_ID_nonpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	1.899	Exp2_Child_ID_nonpair equivalent	Equivalent effect from independent participants
Variability	1.899	Exp2_Child_ID_nonpair intercept - chance	*Justification A* a = condition, a1 = LV, a2 = HV e = adjusted intercept (note, can't use the intercept from within the current model as it is negative.)

AGE COMPARISON

ADULTS vs KIDS

Model	Effect	H1	COMPUTATION	JUSTIFICATION
Exp3_AgeComp_DM	Age group	1.204	Exp2_AgeComp_DM_AOC age group	Equivalent effect from independent participants
Exp3_AgeComp_Train	Age group	0.186	Exp2_AgeComp_Train_AOC age group	Equivalent effect from independent participants
Exp3_AgeComp_ID_minpair	Age group	0.251	Exp2_AgeComp_ID_AOC_minp age group	Equivalent effect from independent participants
Exp3_AgeComp_ID_nonpair	Age group	1.441	Exp2_AgeComp_ID_AOC_nonp age group	Equivalent effect from independent participants

EXPERIMENT COMPARISON 2 vs 3

ExpComp_Adult_Train

Effect	H1	COMPUTATION	JUSTIFICATION
experiment	0.195	ExpComp_Adult_Training intercept - chance	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = adjusted intercept
experiment :block	0.122	ExpComp_Adult_Training block	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = block

ExpComp_Adult_ID

Model	Effect	H1	COMPUTATION	JUSTIFICATION
ExpComp_Adult_ID_minpair	experiment	0.189	ExpComp_Adult_ Identification_minp intercept – chance	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = adjusted intercept
ExpComp_Adult_ID_nonpair	experiment	1.368	ExpComp_Adult_ Identification_nonp intercept – chance	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = adjusted intercept

ExpComp_Child_Train

Effect	H1	COMPUTATION	JUSTIFICATION
experiment	0.195	ExpComp_Adult_Training intercept - chance	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = adjusted intercept
experiment :block	0.122	ExpComp_Adult_Training block	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = block

ExpComp_Child_ID

Model	Effect	H1	COMPUTATION	JUSTIFICATION
ExpComp_Child _ID_minpair	experiment	0.137	ExpComp_Child_ Identification_minp intercept - chance	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = adjusted intercept
ExpComp_Child _ID_nonpair	experiment	0.301	ExpComp_Child_ Identification_nonp intercept - chance	*Justification A* a = experiment, a1 = exp3, a2 = exp2 e = adjusted intercept

STUDY 2

The pre-registration specified that wherever an effect was found in one age group, it would be used to inform the effect size in the other age group. However, it also specified how H1 would be computed if this value was not available. In the tables below, it is indicated how H1 was computed if neither of the options were available, and where computations deviated from the pre-registration.

TRAINING

TrainG4

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	2.362	TrainG8 intercept - chance	Equivalent effect from independent participants
Session	0.166	TrainG8 session	Equivalent effect from independent participants
Condition	0.648	TrainG8 condition	Equivalent effect from independent participants
Condition : session	0.103	Condition:session in child model (Giannakopoulou et al., 2017)	Equivalent effect from independent participants

TrainG8

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.806	TrainG4 intercept - chance	Equivalent effect from independent participants
Session	0.033	TrainG4 session	Equivalent effect from independent participants
Condition	1.626	condition in model with adults from (Giannakopoulou et al., 2017)	Equivalent effect from independent participants
Condition : session	0.103	Condition:session in model with children from (Giannakopoulou et al., 2017)	Equivalent effect from independent participants

TrainAgeComp

Effect	H1	COMPUTATION	JUSTIFICATION
Age	1.998	Age from model with children from (Giannakopoulou et al., 2017)	Equivalent effect from independent participants
Age:session	0.083	TrainAgeComp session	*Justification A* a = age, a1 = 7yo, a2 = 11yo e = session

DISCRIM

DiscrimG4

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	1.689	DiscrimG8 intercept – chance	Equivalent effect from independent participants
Session	0.219	DiscrimG8 session	Equivalent effect from independent participants
Condition : session	0.219	DiscrimG8 session	*Justification A* a = condition, a1 = LV, a2 = HV e = session **PRE-REG NOTE:** Note, we planned to use the effect of session from the current model, but this effect isn't seen in the G4 data.

Condition: session: novelty	0.146	(DiscrimG8 session) * (2/3)	*Justification C* a = novelty, a1 = untrained, a2 = trained b = condition, b1 = LV, b2 = HV e = session **PRE-REG NOTE:** Note, we planned to use the effect of session by condition from the current model but this is not found in the G4 data. The calculation also differs from the pre-registration, as there was an error in the calculation of the effect, which would overestimate it.

DiscrimG8

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.833	DiscrimG4 intercept - chance	Equivalent effect from independent participants
Session	1.689	DiscrimG8 intercept - chance	*Justification A* a = session, a1 =pre, a2 = post e = adjusted intercept
Condition : session	0.219	DiscrimG8 session	*Justification A* a = condition, a1 = LV, a2 = HV e = session
Condition: session: novelty	0.219	(DiscrimG8 session) * (2/3)	*Justification C* a = novelty, a1 = untrained, a2 = trained b = condition, b1 = LV, b2 = HV e = intercept **PRE-REG NOTE:** This differs from the pre-registration, as there was an error in the calculation of the effect, which would overestimate it.

DiscrimAgeComp

Effect	H1	COMPUTATION	JUSTIFICATION
Age	1.164	DiscrimAgeComp intercept - chance	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = adjusted intercept
Age:session	2.33	(DiscrimAgeComp intercept – chance) * (2/3)	*Justification C* a = age, a1 = 7yo, a2 = 1yo b = session, b1 = pre, b2 = post e = intercept **PRE-REG NOTE:** Note, we had planned to use effect of session in the current model but this effect isn't seen in the current data.
Session: age : condition	2.33	DiscrimAgeComp session:age	*Justification A* a = condition, a1 = LV, a2 = HV e = session by age **PRE-REG NOTE:** Note, we had planned to use effect of session in the current model but this effect isn't seen in the current data. Using the effect of session by age was deemed more appropriate than using the adjusted intercept.

ORTHOGRAPHY ID

OrthIDG4

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.978	OrthIDG8 intercept - chance	Equivalent effect from independent participants
Session	0.228	OrthIDG4 intercept – chance	*Justification A* a = session, a1 =pre, a2 = post e = adjusted intercept
Condition : session	0.07	OrthIDG8 session	*Justification A* a = condition, a1 = LV, a2 = HV e = session **PRE-REG NOTE:** Note, we had planned to use effect of session in the current model but this effect isn't seen in the G4 data.

Condition: session: novelty	0.023	(OrthIDG8 session)* (2/3)	*Justification C* a = novelty, a1 = untrained, a2 = trained b = condition, b1 = LV, b2 = HV e = session **PRE-REG NOTE:** Note, we had planned to use effect of block in the current model but this effect isn't seen in the G4 data. The analysis also differs from the pre-registration, as there was an error in the calculation of the effect, which would overestimate it.

OrthIDG8

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.228	OrthIDG4 intercept - chance	Equivalent effect from independent participants
Session	0.978	OrthIDG8 intercept - chance	*Justification A* a = session, a1 =pre, a2 = post e = adjusted intercept
Condition : session	0.035	OrthIDG8 session	*Justification A* a = condition, a1 = LV, a2 = HV e = session
Condition: session: novelty	0.023	(OrthIDG8 session) * (2/3)	*Justification C* a = novelty, a1 = untrained, a2 = trained b = condition, b1 = LV, b2 = HV e = session **PRE-REG NOTE:** This differs from the pre-registration, as there was an error in the calculation of the effect, which would overestimate it.

OrthIDAgeComp

Effect	H1	COMPUTATION	JUSTIFICATION
Age	0.535	OrthIDAgeComp intercept - chance	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = adjusted intercept
Age:session	0.356	(OrthIDAgeComp intercept – chance) * (2/3)	*Justification C* a = session, a1 =pre, a2 = post b = age, b1 = 7yo, b2 = 11yo e = adjusted intercept **PRE-REG NOTE:** This differs from the pre-registration, as there was an error in the calculation of the effect, which would overestimate it.
Session: age : condition	0.713	DiscrimAgeComp session:age	*Justification A* a = condition, a1 = LV, a2 = HV e = session by age **PRE-REG NOTE:** Note, we had planned to use effect of session in the current model but this effect isn't seen in the data. Using the session by age effect was deemed more appropriate than the adjusted intercept.

PICTURE ID

Note that, as mentioned in the thesis, all analyses described here differ from the pre-registration, since that document overlooked the fact that this test does not have item-novelty as a factor. Instead, the within-participant manipulation is whether trials are minimal pair or non-minimal pairs. These types of trials are quite different from each other and comparisons between these types of trial are not as much of interest as whether the variability effect is seen in each case. Therefore, these two types of trial were analysed separately for each age group.

PicIDG4minpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	1.373	PicIDG8minpair intercept - chance.	Equivalent effect from independent participants
Session	0.767	PicIDG8minpair session	Equivalent effect from independent participants
Condition : session	0.224	PicIDG4minpair session	*Justification A* a = condition, a1 = LV, a2 = HV e = session

PicIDG4nonpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	3.277	PicIDG8nonpair intercept - chance.	Equivalent effect from independent participants
Session	0.572	PicIDG4minpair intercept - chance	*Justification A* a = session, a1 =pre, a2 = post e = adjusted intercept
Condition : session	0.381	(PicIDG4minpair intercept – chance) * (2/3)	*Justification C* a = session, a1 =pre, a2 = post b = condition, b1 = LV, b2 = HV e = adjusted intercept

PicIDG8minpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.224	PicIDG4minpair intercept - chance	Equivalent effect from independent participants
Session	0.14	PicIDG4minpair session	Equivalent effect from independent participants
Condition : session	0.767	PicIDG4minpair session	*Justification A* a = condition, a1 = LV, a2 = HV e = session

PicIDG8nonpair

Effect	H1	COMPUTATION	JUSTIFICATION
Chance	0.224	PicIDG4nonpair intercept - chance	Equivalent effect from independent participants
Session	3.277	PicIDG4minpair intercept - chance	*Justification A* a = session, a1 =pre, a2 = post e = adjusted intercept
Condition : session	2.185	(PicIDG4minpair intercept – chance) * (2/3)	*Justification C* a = session, a1 =pre, a2 = post b = condition, b1 = LV, b2 = HV e = adjusted intercept

PicIDAgeCompminpair

Effect	H1	COMPUTATION	JUSTIFICATION
age	0.682	PicIDAgeComp intercept - chance	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = adjusted intercept
Session:age	0.393	PicIDAgeComp session	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = session
Session: age: variability	0.262	DiscrimAgeComp session:age	*Justification A* a = condition, a1 = LV, a2 = HV e = session by age

PicIDAgeCompnonpair

Effect	H1	COMPUTATION	JUSTIFICATION
age	1.752	PicIDAgeComp intercept - chance	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = intercept
Session:age	0.22	PicIDAgeComp session	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = session
Session: age: variability	0.44	PicIDAgeComp session : age	*Justification A* a = condition, a1 = LV, a2 = HV e = session by age

VOCABULARY

VocabG4

Effect	H1	COMPUTATION	JUSTIFICATION
Session	1.006	Value from the equivalent effect in VocabG8 model.	Equivalent effect from independent participants
Condition : session	1.006	VocabG8 session	*Justification A* a = condition, a1 = LV, a2 = HV e = session **PRE-REG NOTE:** Note, we had planned to use effect of session in the current model but this effect isn't seen in the current data.

VocabG8

Effect	H1	COMPUTATION	JUSTIFICATION
Session	6.473	VocabG8 intercept - chance	*Justification A* a = session, a1 =pre, a2 = post e = adjusted intercept
Condition : session	1.006	VocabG8 session	*Justification A* a = condition, a1 = LV, a2 = HV e = session

VocabAgeComp

Effect	H1	COMPUTATION	JUSTIFICATION
age	3.349	VocabAgeComp intercept - chance	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = intercept
Session:age	0.315	VocabAgeComp session	*Justification A* a = group, a1 = 7yo, a2 = 11yo e = session
Session: age: variability	0.21	(VocabAgeComp session) * (2/3)	*Justification C* e = session HV = a2; LV = a1; G4 = b2, G8 = b1 **PRE-REG NOTE:** This differs from the pre-registration, as there was an error in the calculation of the effect, which would overestimate it.

PRODUCTION

ProdG4

Effect	H1	COMPUTATION	JUSTIFICATION
Session	8.14	maximum difference = max NS score - pre-test score = 8.11 - (-0.28) = 8.39 Use half of this for the estimate	Maximum possible score, based on the maximum accuracy prime rating for any of the native speaker control trials, and the score at pre-test.
Condition : session	8.14		
Condition : session:novelty	8.14		

ProdG8

Effect	H1	COMPUTATION	JUSTIFICATION
Session	7.785	maximum difference = max NS score - pre-test score = 8.11 - (0.43) = 7.68 Use half of this for the estimate	Maximum possible score, based on the maximum accuracy prime rating for any of the native speaker control trials, and the score at pre-test.
Condition : session	7.785		
Condition : session:novelty	7.785		

Justifications

A) Estimation of effect *a.e* – i.e. modulation of an effect *e* by a factor *a*, where *a* has two levels, based on $\overline{e}$

- $\overline{e}$ is the grandmean effect of e
- $a_1.e$ is the effect of *e* at level a_1 of a
- $a_2.e$ is the effect of *e* at level a_2 of a
- $a.e = a_2.e - a_1.e$

Assumption for plausible maximum:

- minimum value effect ***e*** in any cell is 0
- $a_1.e = 0$

$a.e = a_2.e - 0$

$\overline{e} = (a_1.e + a_2.e)/2$

$\overline{e} = (0 + a.e)/2$

$2\overline{e} = a.e$

Estimate is set to half the maximum:

estimate = $\overline{e}$

B) Estimation of effect *a.e* – i.e. modulation of an effect *e* by a factor *a*, where *a* is a linear predictor with *n* levels or a continuous predictor of length *n* - 1, based on mean $\overline{e}$

- $\overline{e}$ is the grandmean effect of e
- *n* is the number of levels of *a* (or the length (number of units) + 1 for a continuous factor)
- $a_i.e$ is the effect of *e* at the i^{th} level/unit
- $a.e = a_{i+1}.e - a_i.e$ (i.e. increase in *e* for a one level/per one unit increase in *a*)

In the case n = 2, this is equivalent to a factor – note that in that case: estimate = $\overline{e}(n)$ /((n-1)(n)) = $\overline{e}$, as in Justification A above, which is in fact a special case of the more general case described here in B.

Assumptions for plausible maximum:

- minimal effect of *e* in any cell is 0
- $a_1.e = 0$

$\overline{e}$ = $(a_1.e + a_2.e + a_3.e \ldots + a_n.e)/n$

$\overline{e}$ = $(0 + 1(a.e) + 2(a.e) \ldots + (n\text{-}1)\ a.e)/n$

Using the formula for triangular numbers:

$\overline{e}$ = $((a.e)(n\text{-}1)(n)/2)\ /n$

$\overline{e}(n)$ = $(a.e)(n\text{-}1)(n)/2$

$2\overline{e}(n)$ = $(a.e)(n\text{-}1)(n)$

$a.e$ = $2\overline{e}(n)\ /\ ((n\text{-}1)(n))$

$a.e$ = $2\overline{e}\ /\ (n\text{-}1)$

Estimate is set to half the maximum:

estimate = $\overline{e}/(n\text{-}1)$

C) Estimation of effect *a.b.e* – i.e. modulation of an effect *e* by interaction *a by b,* where each of *a* and *b* are factors with two levels, based on mean $\overline{e}$

- $\overline{e}$ is the grandmean effect of e
- $a_i.\ b_j$ is the effect of *e* in cell $a_i,\ b_j$
- $a.b.e = (a_2.b_2.e - a_2.b_1.e) - (a_1.b_2.e - a_1.b_1.e)$

Assumptions for plausible maximum:

- minimal effect of *e* = 0
- $a_2.b_1.e = 0$
- $a_1.b_2.e = a_1.b_1.e \neq 0$ [i.e. for level a_1, there is no effect of b on e; however we don't assume no effect of e in this condition]
- $a_2.b_2.e = a_1.b_2.e = a_1.b_1.e$ [i.e. in all cells where there is an effect of e, this effect will be the same]

$a.b.e$ = $(a_2.b_2.e - a_2.b_1.e) - (a_1.b_2.e - a_1.b_1.e)$

$a.b.e$ = $(a_2.b_2.e - 0) - (a_2.b_2.e - a_2.b_2.e)$

$a.b.e$ = $a_2.b_2.e$

$\overline{e}$ = $(a_2.b_2.e + a_2.b_1.e + a_1.b_2.e + a_1.b_1.e)/4$

$\overline{e}$ = $(3a_2.b_2.e + 0)/4$

$4\overline{e}$ = $3a_2.b_2.e$

$1⅓\ \overline{e}$ = $a_2.b_2.e$

$a.b.e$ = $1⅓\ \overline{e}$

Estimate is set to half the maximum:

estimate = ⅔ $\overline{e}$

D) Estimation of effect *a.b.c.e* – i.e. modulation of an effect *e* by interaction *a by b by c,* where each of *a b* and c are factors with two levels, based on mean $\overline{e}$

- $\overline{e}$ is the grandmean effect of e
- $a_i.\ b_j.\ c_k$ is the effect of *e* in cell $a_i,\ b_j,\ c_k$
- $a.b.c.e = ((a_2.\ b_2.\ c_2.e - a_2.\ b_2.\ c_1.e) - (a_2.\ b_1.\ c_2.e - a_2.\ b_1.\ c_1.e)) - ((a_1.\ b_2.\ c_2.e - a_1.\ b_2.\ c_1.e) - (a_1.\ b_1.\ c_2.e - a_1.\ b_1.\ c_1.e))$

Assumptions for plausible maximum:

- minimal effect of *e* = 0
- $a_2.\ b_2.\ c_1.e = 0$; $a_2.\ b_1.\ c_2.e = 0$; $a_2.\ b_1.\ C_1 = 0$ (for $a=a_2$, don't have effect of *e*, except where $b=b_2$ and $c=c_2$)
- $(a_1.\ b_2.\ c_2.e - a_1.\ b_2.\ c_1.e) - (a_1.\ b_1.\ c_2.e - a_1.\ b_1.\ c_1.e) = 0$ (for $a=a_1$, effect of *c* on *e* is equivalent for $b=b_1$ and $b=b_2$)
- $a_2.\ b_2.\ c_2.e = a_1.\ b_2.\ c_2.e = a_1.\ b_1.\ c_2.e$ (effect of *e* in condition a_2 where $b=b_2$ and $c=c_2$ equals the effect of *e* in condition a_1 where $c=c_2$)

$a.b.c.e = ((a_2.\ b_2.\ c_2.e - a_2.\ b_2.\ c_1.e) - (a_2.\ b_1.\ c_2.e - a_2.\ b_1.\ c_1.e)) - ((a_1.\ b_2.\ c_2.e - a_1.\ b_2.\ c_1.e) - (a_1.\ b_1.\ c_2.e - a_1.\ b_1.\ c_1.e))$

$a.b.c.e = ((a_2.\ b_2.\ c_2.e - 0) - (0 - 0)) - 0$

$a.b.c.e = a_2.\ b_2.\ c_2.e$

$\overline{e} = (a_2.\ b_2.\ c_2.e + a_2.\ b_2.\ c_1.e + a_2.\ b_1.\ c_2.e + a_2.\ b_1.\ c_1.e$

$+ a_1.\ b_2.\ c_2.e + a_1.\ b_2.\ c_1.e + a_1.\ b_1.\ c_2.e + a_1.\ b_1.\ c_1.e) / 8$

$\overline{e} = (a_2.\ b_2.\ c_2.e + 0 + 0 + 0 + a_2.\ b_2.\ c_2.e + 0 + a_2.\ b_2.\ c_2.e + 0) / 8$

$8\overline{e} = 3a_2.\ b_2.\ c_2.e$

$8\overline{e} = 3a.b.c.e$

$a.b.c.e = 2⅔\overline{e}$

Estimate is set to half the maximum:

estimate = $1⅓\overline{e}$

Appendix IV. Analyses Study 2 without FLEECE-THOUGHT control contrast

Training

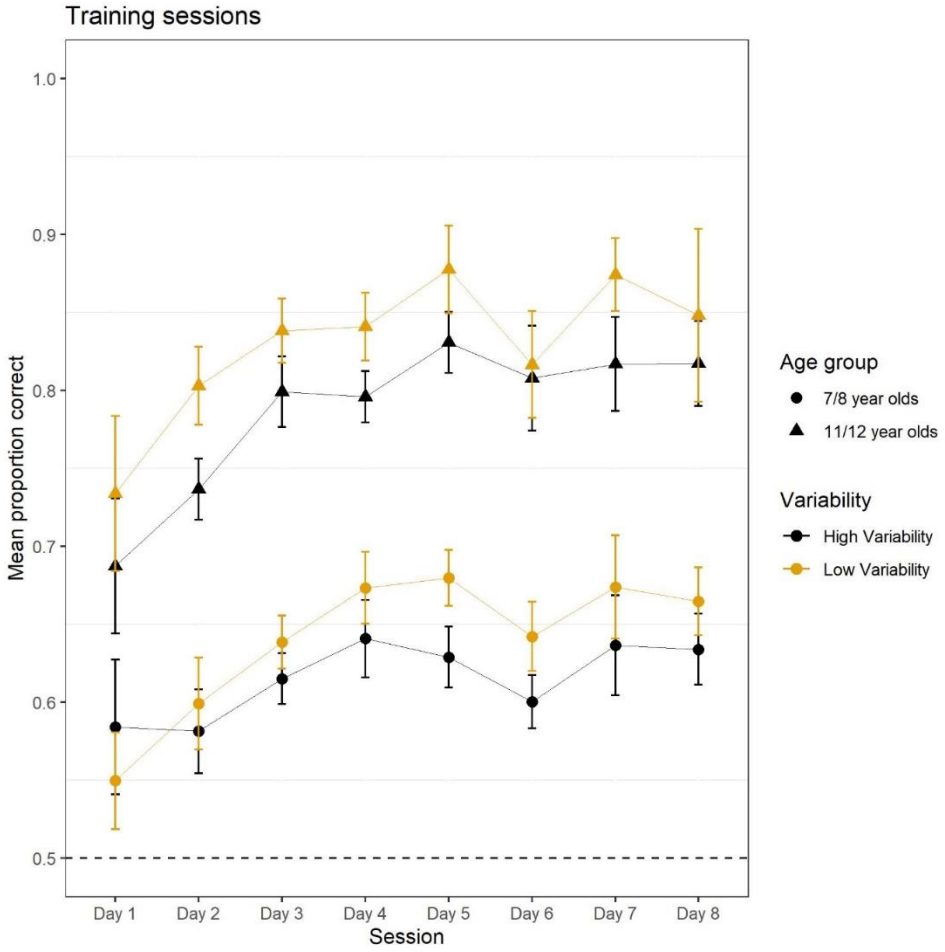


Figure 54. Accuracy results for 7-8 year olds and 11-12 year olds during Training of Experiment 5 without the FLEECE-THOUGHT contrast in, comparing accuracy for HV versus LV training input. The error bars indicate 95% CI, and the dashed line indicates chance level.

Final structure TrainG4 model: accuracy ~ session*condition + VowelContrast + (session:condition|participant) + (1|item)

Hypothesis	fixed effect in model	β	SE	z	P	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.572	0.093	6.128	<.001	2.362	$1.09*10^{7}$	[0.0156, >5.5156]
improvement from pre- to post- test	session	0.042	0.009	4.545	<.001	0.166	3267.28	[0, >5.5156]
LV outperforming HV	condition	0.144	0.131	1.100	.271	0.648	0.606	[0, 1.2156]
greater overall improvement for LV training	Interaction condition: session	0.029	0.019	1.565	.118	0.103	1.09	[0, 0.3156]

Table 73. Mixed model results for the Training analysis without the FLEECE-THOUGHT contrast, for 7-8 year olds.

Final structure TrainG8 model: accuracy ~ session*condition + VowelContrast + talker + (session*condition|participant) + (session:condition||item)

Hypothesis	fixed effect in model	beta	SE	z	P	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	1.892	0.198	9.531	<.001	0.806	$1.89*10^{18}$	[0.1156, >5.5156]
improvement from pre- to post- test	session	0.194	0.026	7.413	<.001	0.033	$2.63*10^{7}$	[0.0156, >5.5156]
LV outperforming HV	condition	0.464	0.235	1.973	.048	1.626	1.881	[1.0156, >5.5156]
greater overall improvement for LV training	Interaction condition: session	0.057	0.070	0.807	.420	0.103	1.051	[0.0156, 0.4156]

Table 74. Mixed model results for the Training analysis without the FLEECE-THOUGHT contrast, for 11-12 year olds.

Discrimination

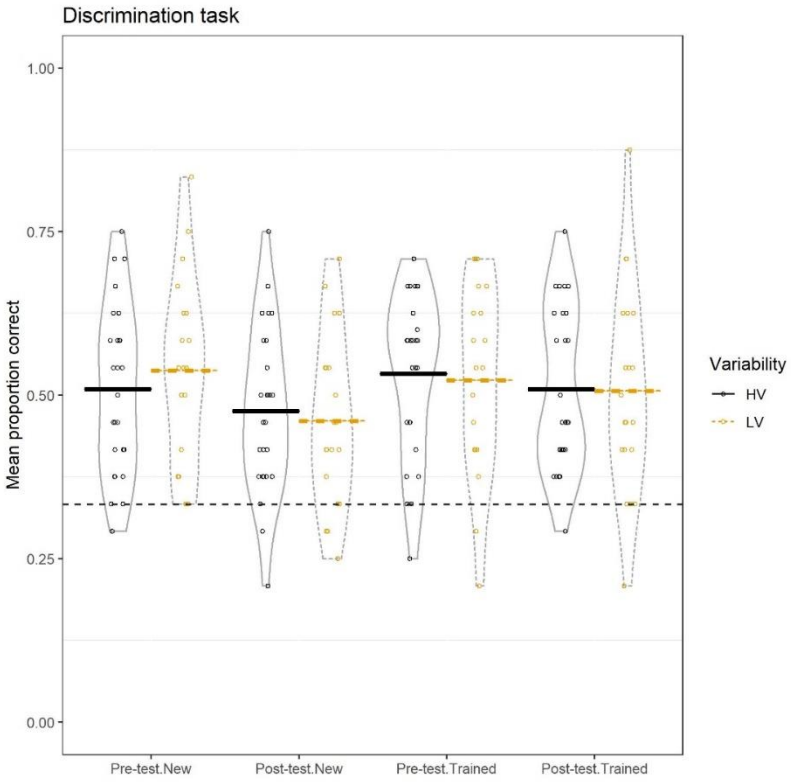


Figure 55. Accuracy results for 7-8 year olds on the pre- and post-test Discrimination task of Experiment 5 without the FLEECE-THOUGHT contrast, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Final structure DiscrimG4 model: accuracy ~ session*condition*itemnovelty + VowelContrast + (session:condition:itemnovelty|participant) + (itemnovelty|item)

Hypothesis	fixed effect in model	beta	SE	z	P	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.717	0.104	6.911	<.001	1.689	$2.63*10^9$	[0, >4.8299]
improvement from pre- to post- test	session	-0.157	0.063	-2.475	0.013	0.219	0.08	[0.1299, >4.8299]
greater overall improvement for HV training	Interaction condition: session	0.090	0.128	0.701	0.483	0.219	0.88	[0.7299, >4.8299]

greater overall improvement for LV training	Interaction condition: session	-0.090	0.128	0.701	0.483	0.219	0.328	[0.2299, >4.8299]
greater overall improvement for HV training for novel items, no diff for trained	Interaction Condition: session: novelty	-0.219	0.256	-0.853	0.394	0.146	0.641	[0, 0.3299]

Table 75. Mixed model results for the Discrimination task without the FLEECE-THOUGHT contrast, for 7-8 year olds.

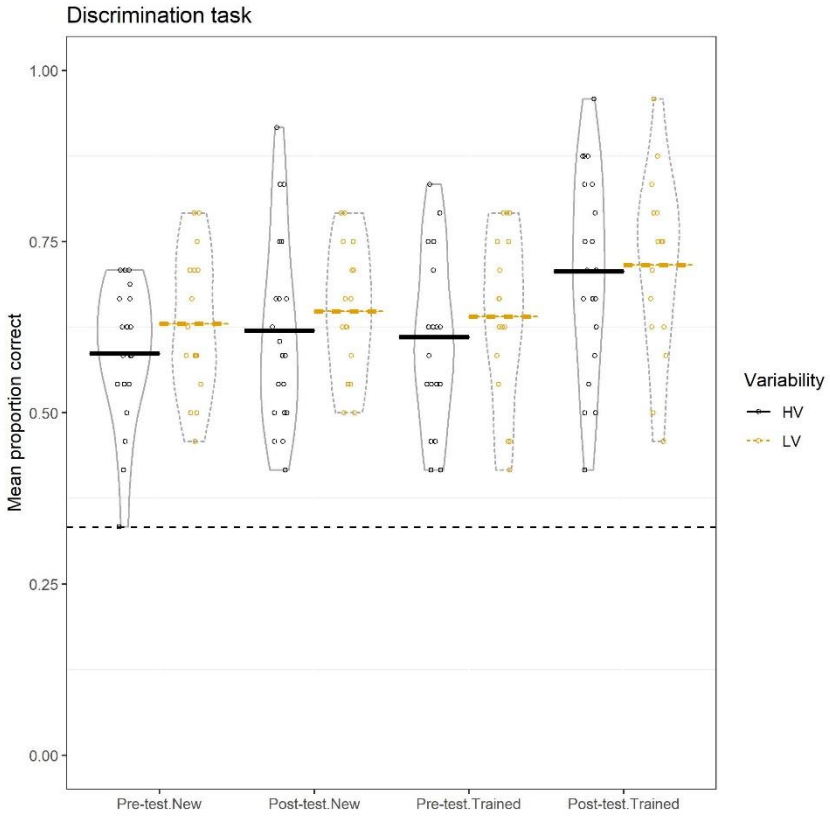


Figure 56. Accuracy results for 11-12 year-olds on the pre- and post-test Discrimination task of Experiment 5 without the FLEECE-THOUGHT contrast, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Final structure DiscrimG8 model: accuracy ~ session *condition*itemnovelty + VowelContrast + (session:condition:itemnovelty|participant) + (itemnovelty|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	1.373	0.130	10.592	<.001	0.717	$1.38*10^{23}$	[0, >4.8299]
improvement from pre- to post- test	session	0.286	0.076	3.788	<.001	1.373	140.497	[0, >4.8299]
greater overall improvement for HV	Interaction condition: session	0.075	0.152	0.494	0.622	0.286	0.693	[0.0299, 0.6299]
greater overall improvement for LV	Interaction condition: session	-0.075	0.152	0.494	0.622	0.286	0.345	[0.0299, 0.2299]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	0.016	0.312	0.051	0.959	0.191	0.8374	[0.0299, 0.8299]

Table 76. Mixed model results for the Discrimination task without the FLEECE-THOUGHT contrast, for 11-12 year olds.

Orthography Identification

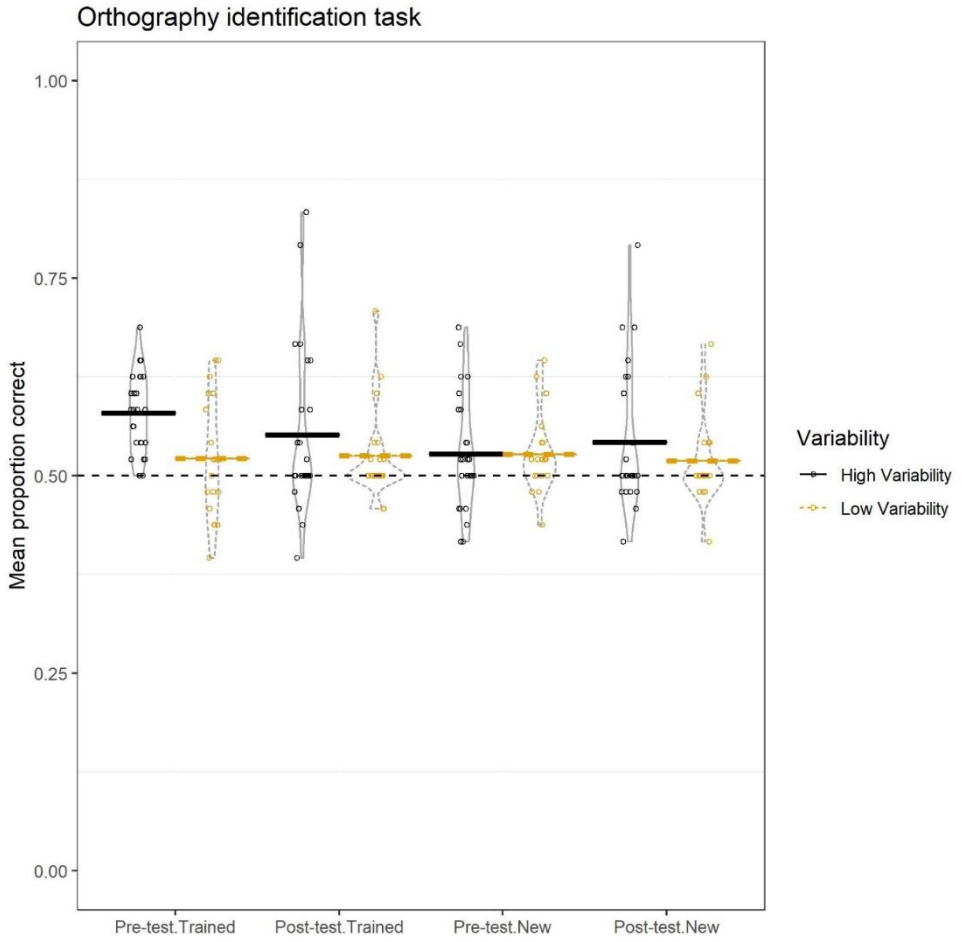


Figure 57. Accuracy results for 7-8 year olds on the pre- and post-test Orthography Identification task of Experiment 5 without the FLEECE-THOUGHT contrast, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Final structure OrthIDG4 model: accuracy ~ session*condition*itemnovelty + VowelContrast + Talker + (session:condition:itemnovelty|participant) + (condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.157	0.034	4.584	<.001	0.978	2522.804	[0, >4.84]
improvement from pre- to post- test	session	-0.019	0.043	-0.440	0.660	0.228	0.198	[0.09, >4.84]

greater overall improvement for HV	Interaction condition: session	-0.015	0.087	-0.177	0.859	0.07	0.717	[0, 0.21]
greater overall improvement for LV	Interaction condition: session	0.015	0.087	-0.177	0.859	0.07	0.856	[0, 0.28]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	-0.219	0.176	-1.239	0.215	0.023	0.877	[0, 0.22]

Table 77. Mixed model results for the Orthography Identification analysis without the FLEECE-THOUGHT contrast, for 7-8 year olds.

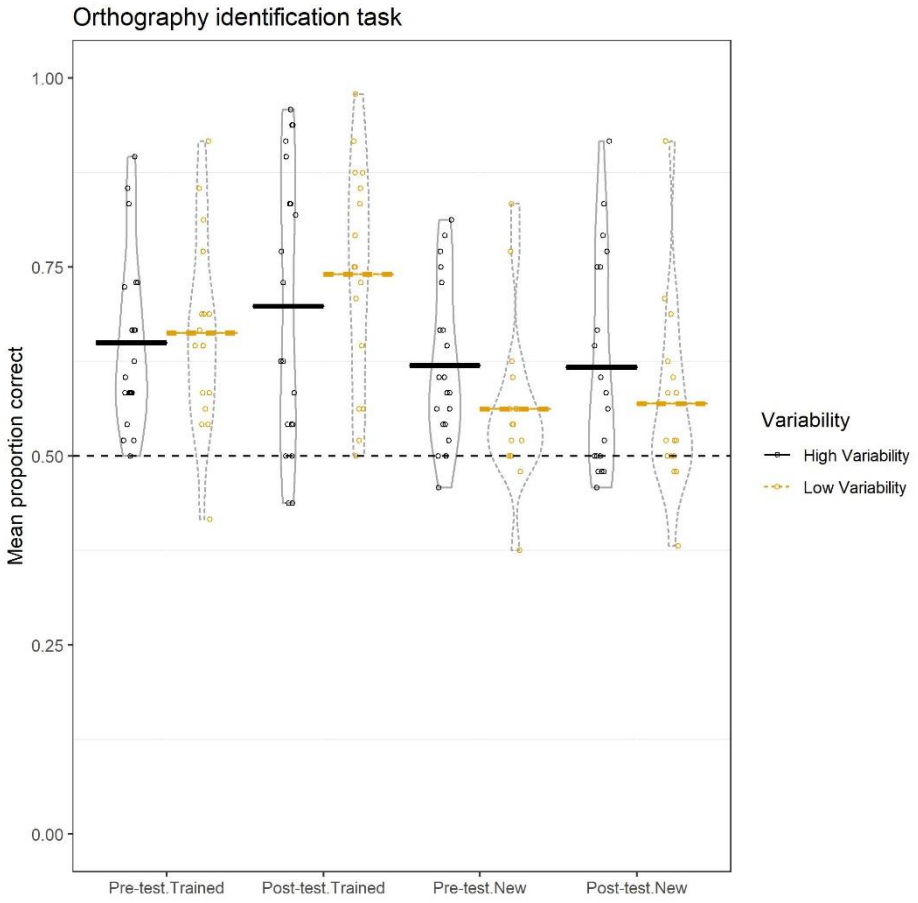


Figure 58. Accuracy results for 11-12 year-olds on the pre- and post-test Orthography Identification task of Experiment 5 without the FLEECE-THOUGHT contrast, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Final structure OrthIDG8 model: accuracy ~ session*condition*itemnovelty + VowelContrast + Talker + (session:condition:itemnovelty|participant) + (session:condition:itemnovelty|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.668	0.109	6.125	<.001	0.228	$3.68*10^6$	[0, >4.84]
improvement from pre- to post- test	session	0.169	0.052	3.237	0.001	0.668	28.355	[0, >4.84]
greater overall improvement for HV	Interaction condition: session	-0.104	0.108	-0.959	0.338	0.169	0.312	[0.16, >4.84]
greater overall improvement for LV	Interaction condition: session	0.104	0.108	0.959	0.338	0.169	1.181	[0, 0.84]
greater overall improvement for HV for novel items, no diff for trained	Interaction Condition: session: novelty	-0.150	0.219	-0.685	0.493	0.112	0.707	[0, 0.37]

Table 78. Mixed model results for the Orthography Identification analysis without the FLEECE-THOUGHT contrast, for 11-12 year olds.

Picture Identification

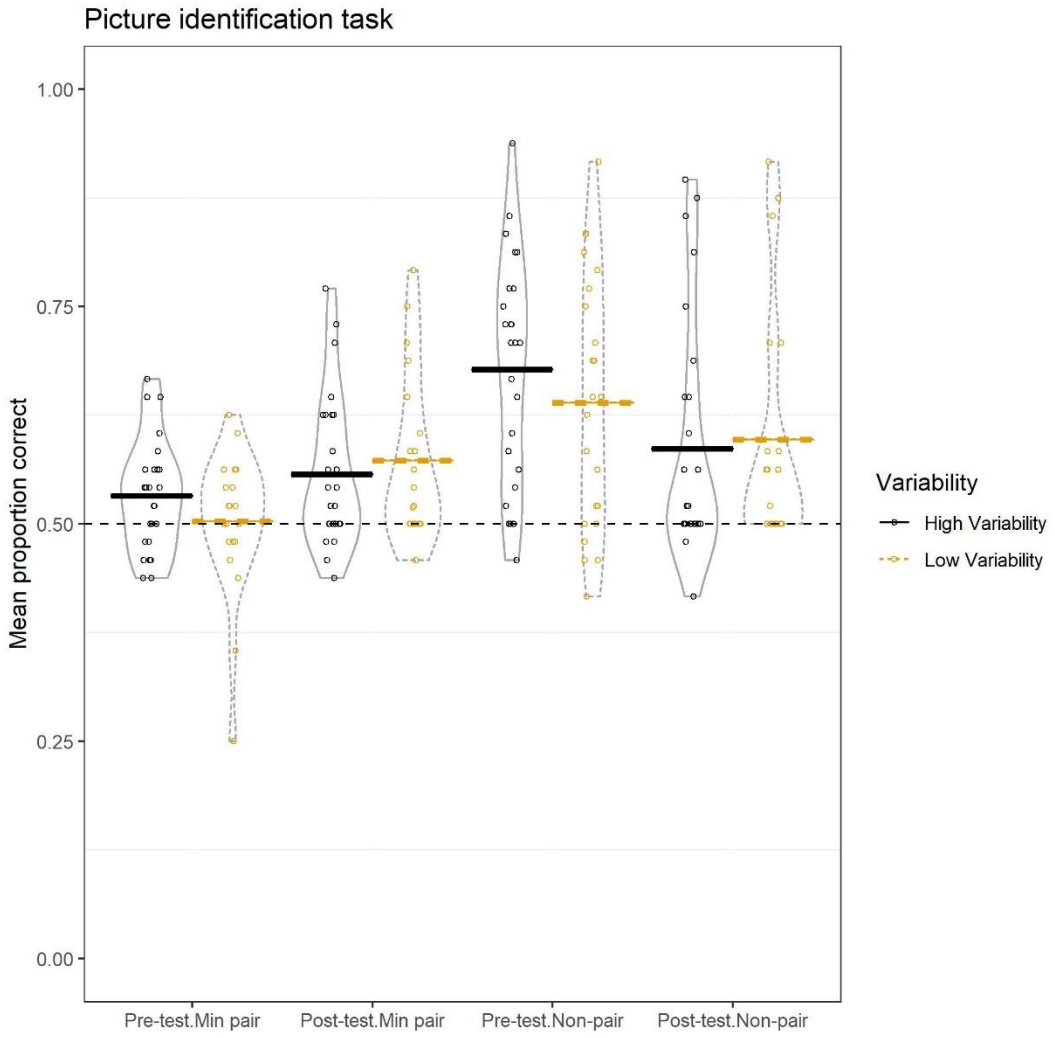


Figure 59. Accuracy results for 7-8 year olds on the pre- and post-test Picture Identification task of Study 2 without the FLEECE-THOUGHT contrast, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Final structure PicIDG4minpair: accuracy ~ session*condition + VowelContrast + Talker + (session*condition||participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.169	0.052	3.246	.001	1.373	14.578	[0, >4.1425]
improvement from pre- to post- test	session	0.183	0.075	2.440	0.015	0.767	3.687	[0.9625, 4.1425]
greater overall improvement for HV	Interaction condition: session	-0.187	0.152	-1.231	0.218	0.169	0.364	[0, 0.1825]
greater overall improvement for LV	Interaction condition: session	0.187	0.152	1.231	0.218	0.169	1.665	[0, 1.7225]

Table 79. Mixed model results for the minimal pair trials in the Picture Identification analysis without the FLEECE-THOUGHT contrast, for 7-8 year olds.

Final structure PicIDG4nonpair: accuracy ~ session*condition + VowelContrast + Talker + (session*condition||participant) + (session:condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.564	0.075	7.491	<.001	3.277	$6.93*10^{10}$	[0, >4.1425]
improvement from pre- to post- test	session	-0.307	0.113	-2.707	.007	0.564	0.055	[0.0825, >4.1425]
greater overall improvement for HV	Interaction condition: session	-0.215	0.231	-0.933	0.351	0.376	0.308	[0.3525, >4.1425]
greater overall improvement for LV	Interaction condition: session	0.215	0.231	0.933	0.351	0.376	1.132	[0, 1.7225]

Table 80. Mixed model results for the non-pair trials in the Picture Identification analysis without the FLEECE-THOUGHT contrast, for 7-8 year-olds.

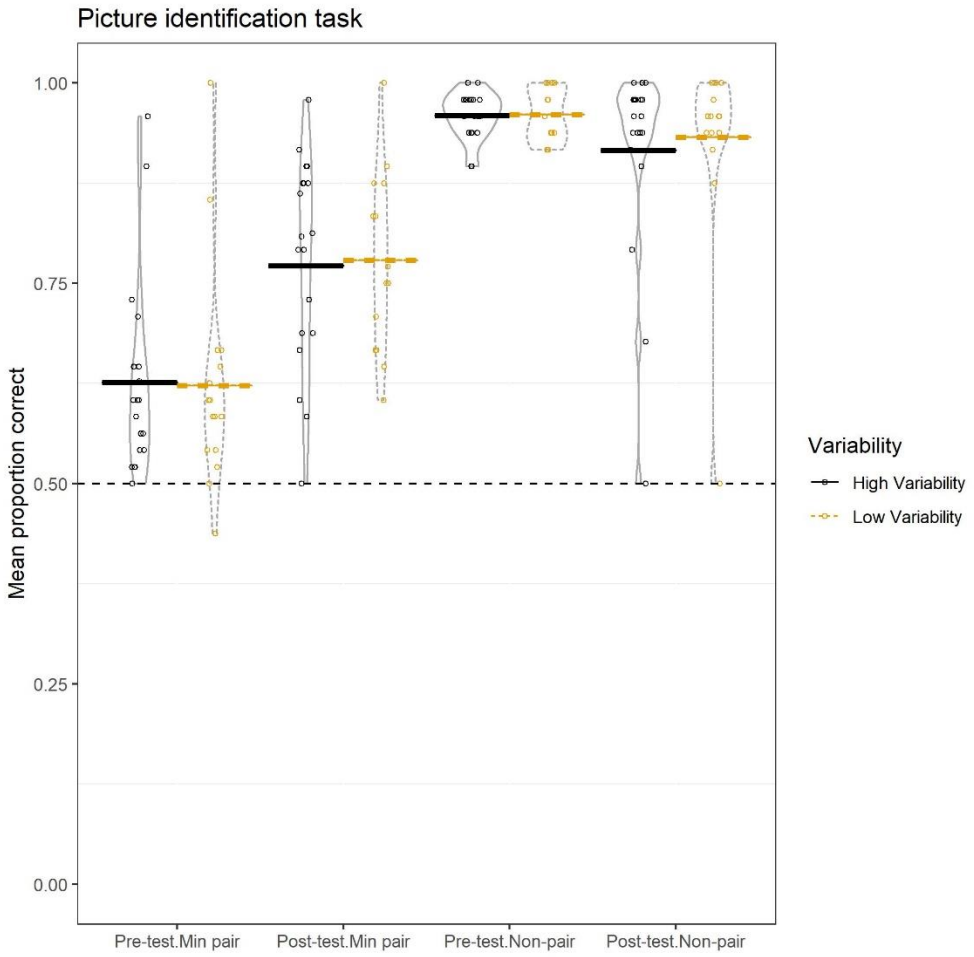


Figure 60. Accuracy results for 11-12 year-olds on the pre- and post-test Picture Identification task of Study 2 without the FLEECE-THOUGHT contrast, comparing accuracy for new versus trained items as well as HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI, and the dashed line indicates chance level.

Final structure PicIDG8minpair: accuracy ~ session*condition + VowelContrast + (session*condition|participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	0.996	0.143	6.955	<.001	0.169	$1.67*10^{6}$	[0, >4.1425]
improvement from pre- to post- test	session	0.840	0.115	7.283	<.001	0.183	$1.88*10^{8}$	[0, >4.1425]
greater overall improvement for HV	Interaction condition: session	-0.005	0.221	-0.023	0.982	0.84	0.248	[0.6225, >4.1425]
greater overall improvement for LV	Interaction condition: session	0.005	0.221	0.023	0.982	0.84	0.257	[0.6425, 4.1425]

Table 81. Mixed model results for the minimal pair trials in the Picture Identification analysis without the FLEECE-THOUGHT contrast, for 11-12 year-olds.

Final structure PicIDG8nonpair: accuracy ~ session*condition + VowelContrast + (session*condition|participant) + (1|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
above chance performance	intercept-chance	3.254	0.166	19.606	<.001	0.169	$1.72*10^{31}$	[0, >4.1425]
improvement from pre- to post- test	session	-0.071	0.279	-0.255	0.798	3.254	0.069	[0.6525, >4.1425]
greater overall improvement for HV	Interaction condition: session	-0.284	0.550	-0.516	0.606	2.169	0.174	[1.0525, >4.1425]
greater overall improvement for LV	Interaction condition: session	0.284	0.550	0.516	0.606	2.169	0.387	[2.5525, >4.1425

Table 82. Mixed model results for the non-pair trials in the Picture Identification analysis without the FLEECE-THOUGHT contrast, for 11-12 year-olds.

Vocabulary

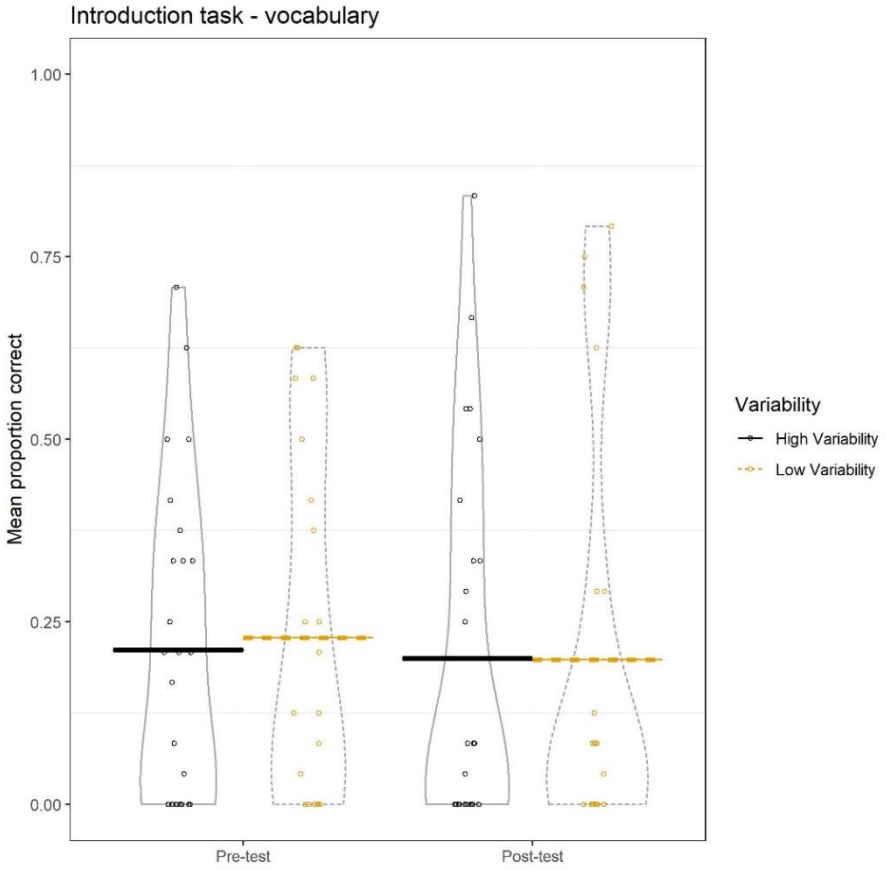


Figure 61. Accuracy results for 7-8 year olds on the Vocabulary task at pre- and post-test without the FLEECE-THOUGHT contrast, comparing accuracy for HV versus LV training input.

Final structure VocabG4 model: accuracy ~ session*condition + Vcontrast + Talker + (session*condition|participant) + (session+condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect *x*	Bayes factor	Robustness Region
improvement from pre- to post- test	session	-0.689	0.416	-1.657	0.098	1.006	0.153	[0, >6.8659]
greater overall improvement for HV	Interaction condition: session	-0.024	0.796	-0.030	0.976	1.006	0.607	[0, 2.1759]
greater overall improvement for LV	Interaction condition: session	0.024	0.796	0.030	0.976	1.006	0.630	[0, 2.2859]

Table 83. Mixed model results for the Vocabulary task for 7-8 year-olds without the FLEECE-THOUGHT contrast.

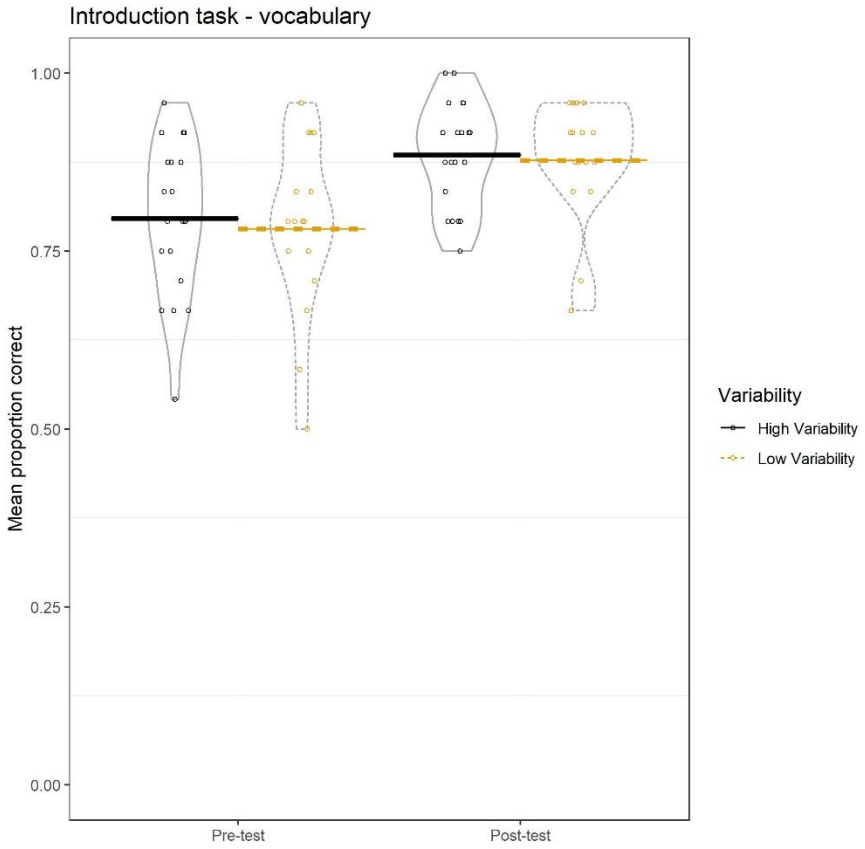


Figure 62. Accuracy results for 11-12 year-olds on the Vocabulary task at pre- and post-test without the FLEECE-THOUGHT contrast, comparing accuracy for HV versus LV training input. The horizontal line in each violin indicates the mean performance with the band around it showing 95% CI. There is no chance level.

Final structure VocabG8 model: accuracy ~ session*condition + Vcontrast + Talker + (session*condition|participant) + (condition|item)

Hypothesis	fixed effect in model	beta	SE	z	p	predicted effect x	Bayes factor	Robustness Region
improvement from pre- to post- test	session	1.046	0.176	5.933	<.001	6.235	$2.45*10^6$	[0.7259, >6.8659]
greater overall improvement for HV	Interaction condition: session	-0.082	0.354	-0.232	0.816	1.046	0.269	[0.8459, >6.8659]
greater overall improvement for LV	Interaction condition: session	0.082	0.354	0.232	0.816	1.046	0.383	[0, 1.2359]

Table 84. Mixed model results for the Vocabulary task without the FLEECE-THOUGHT contrast, for 11-12 year olds.

Appendix V. Study 2 - Language background questionnaire

Original Dutch questionnaire, with English translation in italics.

1. ID Kind - *Child ID*

2. Geboortedatum - *Date of birth*

3. Gender - *Gender*

4. Dyslectisch / Gehoorproblemen / Andere taalproblemen?
   Dyslexia / Hearing problems / DLD
   Evt. meer info - *further details*

5. Welke talen spreek je? - *What languages do you speak?*
   *W*anneer geleerd? - *Learnt when?*

6. Spreekt thuis Nederlands? Zo nee, spreekt:
   Speaks Dutch at home? *If not, speaks:*

7. Ouders/verzorgers die NIET NL als moedertaal hebben?
   Parents/carers who do NOT speak Dutch as L1?

   Spreken deze taal tegen kind?
   Do they speak this language to the child?

OUDER/VERZORGER– *PARENT/CARER*

(provided for up to 4 parents/carers)

Spreekt NL met kind?	Nooit	Af en toe	Soms	Regelmatig	Altijd
Speaks Dutch with child?	*Never*	*Now and then*	*Sometimes*	*Often*	*Always*
Spreekt taal 1 (_) met kind?	*Nooit*	*Af en toe*	*Soms*	*Regelmatig*	*Altijd*
Spreekt taal 2 (_) met jou?	*Nooit*	*Af en toe*	*Soms*	*Regelmatig*	*Altijd*
Speaks Language 1 with child	*Never*	*Now and then*	*Sometimes*	*Often*	*Always*
Speaks Language 2 with child	*Never*	*Now and then*	*Sometimes*	*Often*	*Always*

BROERS/ZUSSEN - SIBLINGS

Spreekt NL met kind? Nooit Af en toe Soms Regelmatig Altijd
Speaks Dutch with child? Never Now and then Sometimes Often Always

Spreekt taal 1 (_) met kind? Nooit Af en toe Soms Regelmatig Altijd
Spreekt taal 2 (_) met jou? Nooit Af en toe Soms Regelmatig Altijd

Speaks Language 1 with child Never Now and then Sometimes Often Always
Speaks Language 2 with child Never Now and then Sometimes Often Always

Hoe vaak televisie/films in EN? Nooit Af en toe Soms Regelmatig Altijd
Watch TV/films in English? Never Now and then Sometimes Often Always

Hoe vaak computerspelletjes in EN? Nooit Af en toe Soms Regelmatig Altijd
Play computer games in English? Never Now and then Sometimes Often Always

Hoe vaak EN muziek/radio? Nooit Af en toe Soms Regelmatig Altijd
Music/radio in English? Never Now and then Sometimes Often Always

Hoe vaak EN boeken/verhalen? Nooit Af en toe Soms Regelmatig Altijd
Books/stories in English? Never Now and then Sometimes Often Always

Wat is de langste opeenvolgende periode ooit in een Engels-sprekend land?
What is the longest continuous period spent in an English-speaking country?

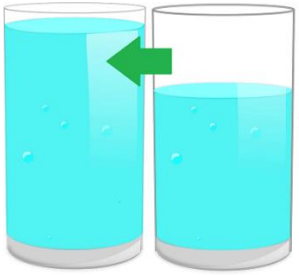
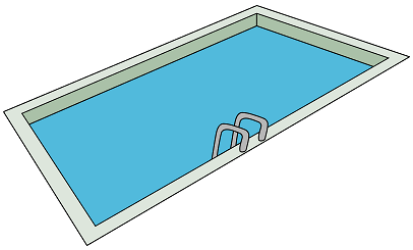
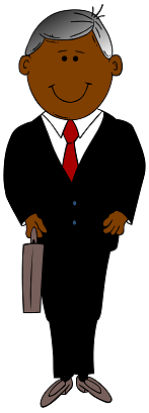
Zijn er (naast school) nog andere situaties waarin je Engels hoort/leest/spreekt?

Are there any situations (outside of school) where you speak/hear/read English?

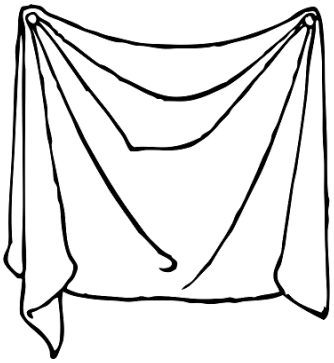
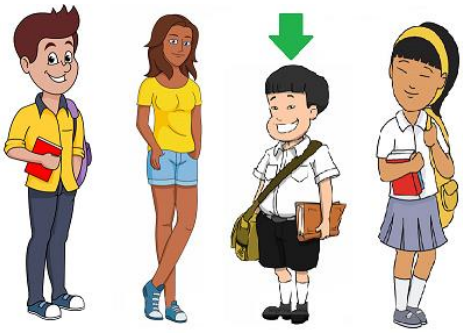
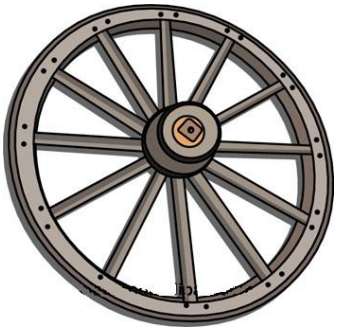
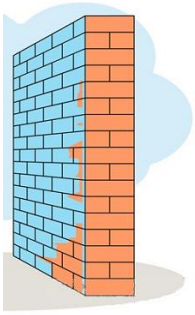
Opmerkingen

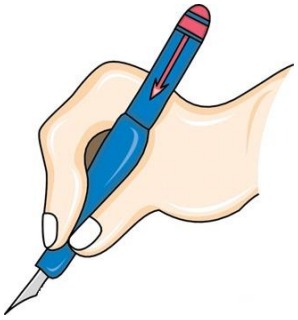
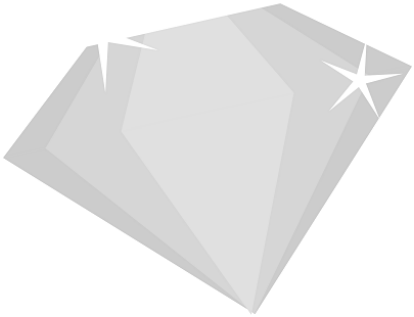
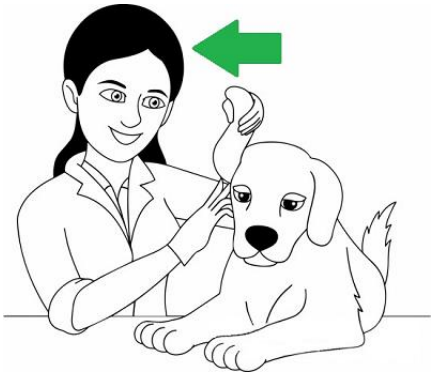
Further comments

Appendix VI. Study 2 - Stimuli pictures

GOOSE- FOOT	
Fool 	Full 
Pool 	Pull 
Suit 	Soot 
Luke 	Look 

STRUT-LOT	
Luck 	Lock 
Cut 	Cot 
Bus 	Boss 
Shut 	Shot 

FLEECE-THOUGHT	
Sheet 	Short 
Wheel 	Wall 
Week 	Walk 
Heel 	Hall 

DRESS-TRAP	
Bed 	Bad 
Pen 	Pan 
Gem 	Jam 
Vet 	Vat 

Appendix VII. Study 2 - Category Boundary task

Stimuli

Stimuli for the Category boundary task consisted of synthesised DRESS-TRAP and STRUT-LOT continua, moving between <gem> - <jam> and <shut> - <shot> respectively. GOOSE-FOOT was not used as the difference between these two vowels is partly one of vowel length rather than vowel quality alone, and FLEECE-THOUGHT was not used as this contrast has end points on opposite ends of the vowel space, meaning that the synthesised continuum would possibly pass through parts of the vowel space that could be labelled as different vowels altogether.

The synthetic speech continua were created using a script written by Mauricio Figueroa (personal communication). This script relies on natural target word recordings to use as continuum endpoints. These target words were recorded by speaker F3, a female native speaker of Standard Southern British English (see Table 85 for the F1 and F2 values of the target vowels used). For each target word, the best recorded exemplar was selected as the reference for the continuum endpoint, before it was segmented and phoneme boundaries within the CVC word were marked using Praat (Boersma & Weenink, 2015). Acoustic models for the vowels were built by extracting formants and formant bandwidth measurements for F1 to F3, as well as pitch and intensity for each endpoint vowel. The total duration for

each continuum pair was set to that of the shortest recorded vowel of the pair (/æ/ and /ʌ/), and each model consists of 200 samples equally distributed across this time. KlattGrid objects were created and populated with the continuum endpoints and 60 steps in between, which were equally distributed across the specified stimulus length. The steps were synthesised to sounds using Klatt synthesis (Klatt & Klatt, 1990), before being spliced back into the end point recordings with an overlap of 10 ms. Mean intensity for the stimuli was scaled to 70 dB. Continua were piloted with several native speakers of English to ensure they sounded reasonably natural.

Vowel	F1	F2
/e/	969 Hz	1798 Hz
/æ/	773 Hz	1676 Hz
/ʌ/	737 Hz	1503 Hz
/ɒ/	583 Hz	1131 Hz

Table 85. Formant values for F1 and F2 for each of the target vowels in the natural recordings from speaker F3, used as reference for the continua endpoints.

Procedure

In the category boundary task, stimuli were presented using an adaptive procedure as described in Hazan, Messaoud-Galusi, Rosen, Nouwens, & Shakespeare (2009); McCarthy, Mahon, Rosen, & Evans (2014); Ramus et al. (2003). Participants were played stimuli from synthesised vowel continua, and were asked to indicate which

of two vowel options they heard by clicking the respective picture on the screen (see Figure 63). Two independent tracks starting from the endpoints of the vowel continua were used (e.g. one at <gem> and one at <jam>), and randomly interleaved throughout the task. Each of the tracks was designed to find the point in their track where trials were identified as belonging to their respective endpoint of that track 71% (and therefore being identified as the other endpoint 29% of the time), based on an adapted two-down/up-up procedure (Baker & Rosen, 2001; Levitt, 1971). Using this adaptive procedure meant participant's categorisation of the vowel continua could be tracked efficiently without the need to test all 60 individual steps of each continuum, ensuring the majority of trials was focussed around the region crucial for estimating their phoneme boundary. When a participant labelled two trials in a row as belonging to the vowel endpoint category that the track started from, the next trial would move further along the vowel continuum and therefore closer towards the region of interest. When a participant then labelled a trial as belonging to the other category than where the track started from, the trial after that would move back towards the endpoint of the continuum (and the starting point of the track) to ensure it would be more likely to be identified as an instance of the original category, thereby focussing in on the category boundary. The initial step size was 10 units of Hz change, reducing linearly to 3 units over the first three reversals. To ensure stable phoneme boundaries were maintained and response consistency was achieved, catch trials presenting the continuum endpoints were

interspersed randomly throughout the task, making up 20% of the total trials. The task ended after 7 reversals, or a maximum of 40 trials. 4 practice trials presenting the participants with the continuum endpoints and providing feedback on their responses were included to familiarise them with the task.

To determine the location of the phoneme category boundary, for each listener their responses to all trials for the 3 vowel continua were aggregated, and a logistic regression was used to obtain a best fit sigmoid function. This provided estimates of the categorisation slope as well as the location of the phoneme boundary. The boundary location was defined as the point on the continuum where a trial was equally likely to be labelled as either endpoint of the continuum and the percept changes from one category to the other. The slope of the identification function reflects the participant's sensitivity to variations in the acoustic feature, whereby a shallower slope indicates the participant is less consistent in labelling the continuum, and thus has less of a categorical perception.

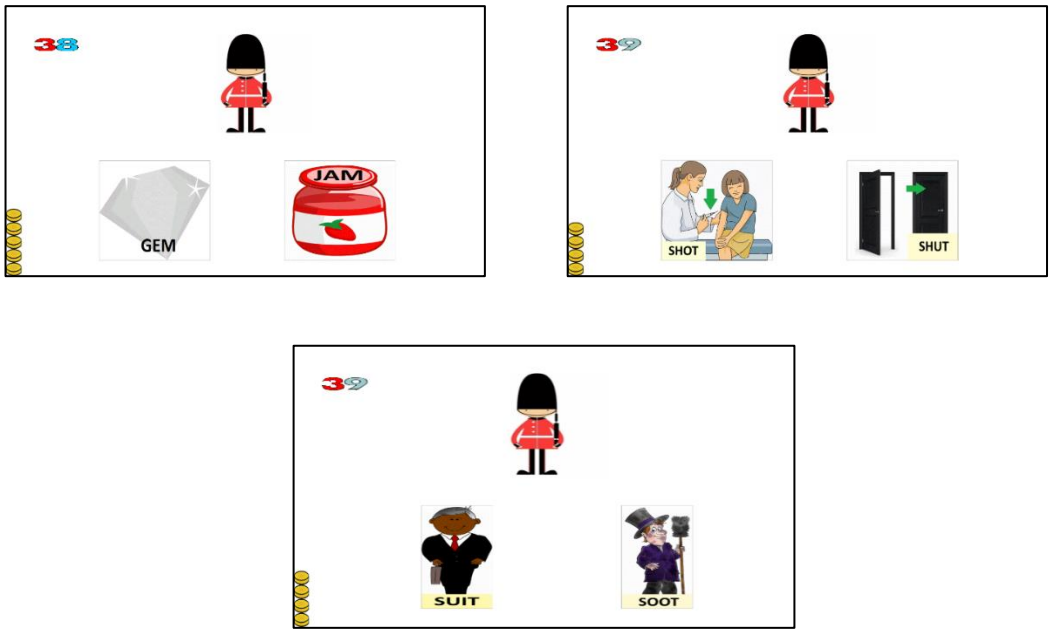


Figure 63. Trial examples for each of the vowel contrasts for the category boundary task.

Results

To exemplify what results from the category boundary task might have looked like, in Figure 64 below presents representative category boundary results from the pilot data performed for Study 2. Top left shows the native SSBE boundary, while top right shows the author of this thesis, an advanced Dutch adult learner of English. The bottom left shows the category boundary for one of the native Dutch 16-year-olds who participated in the pilot test. The bottom right shows the best boundary result of the native Dutch 7-year-olds who participated in the pilot test (note that most had a flat horizontal line, indicating they could not clearly determine a boundary).

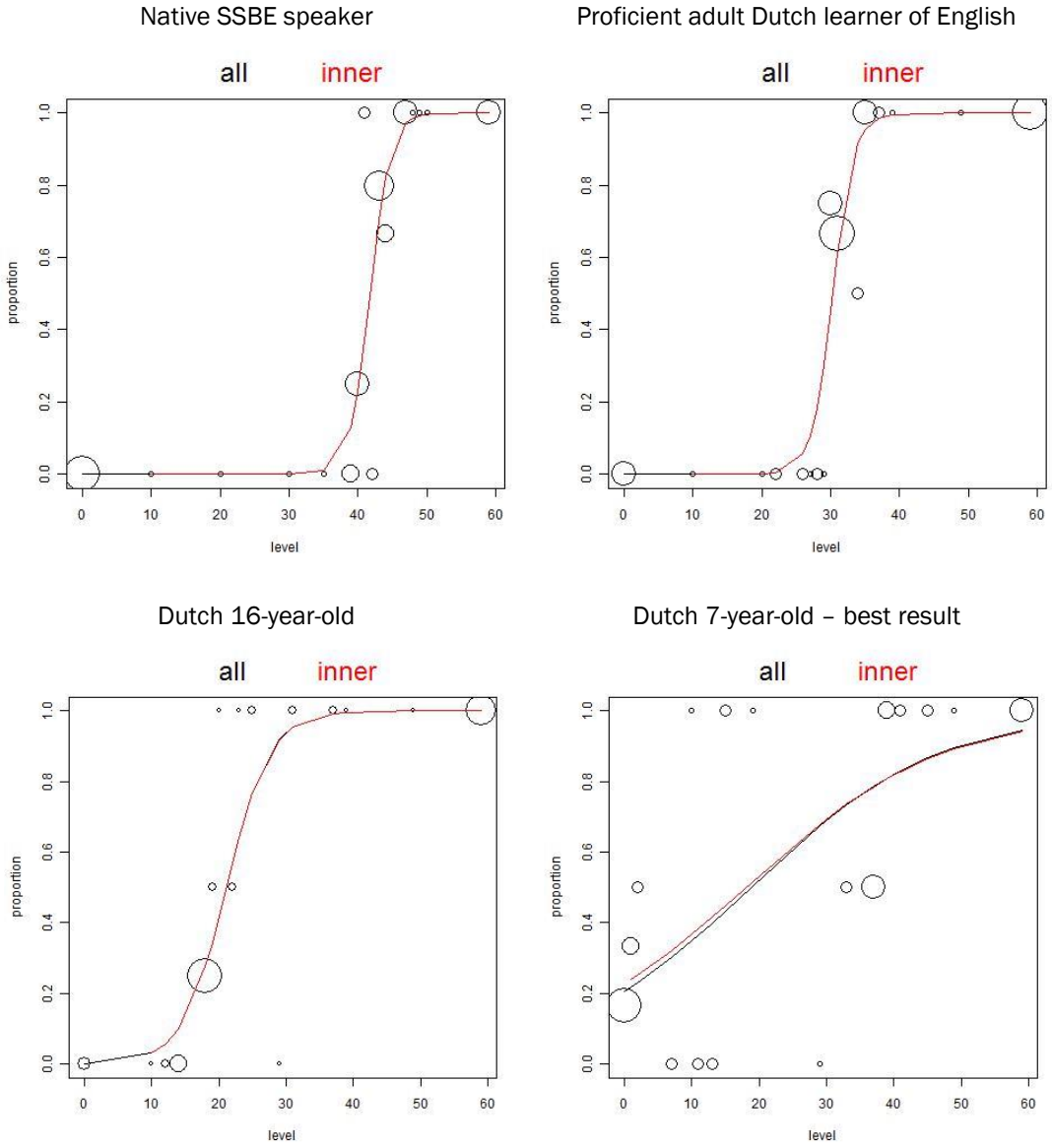
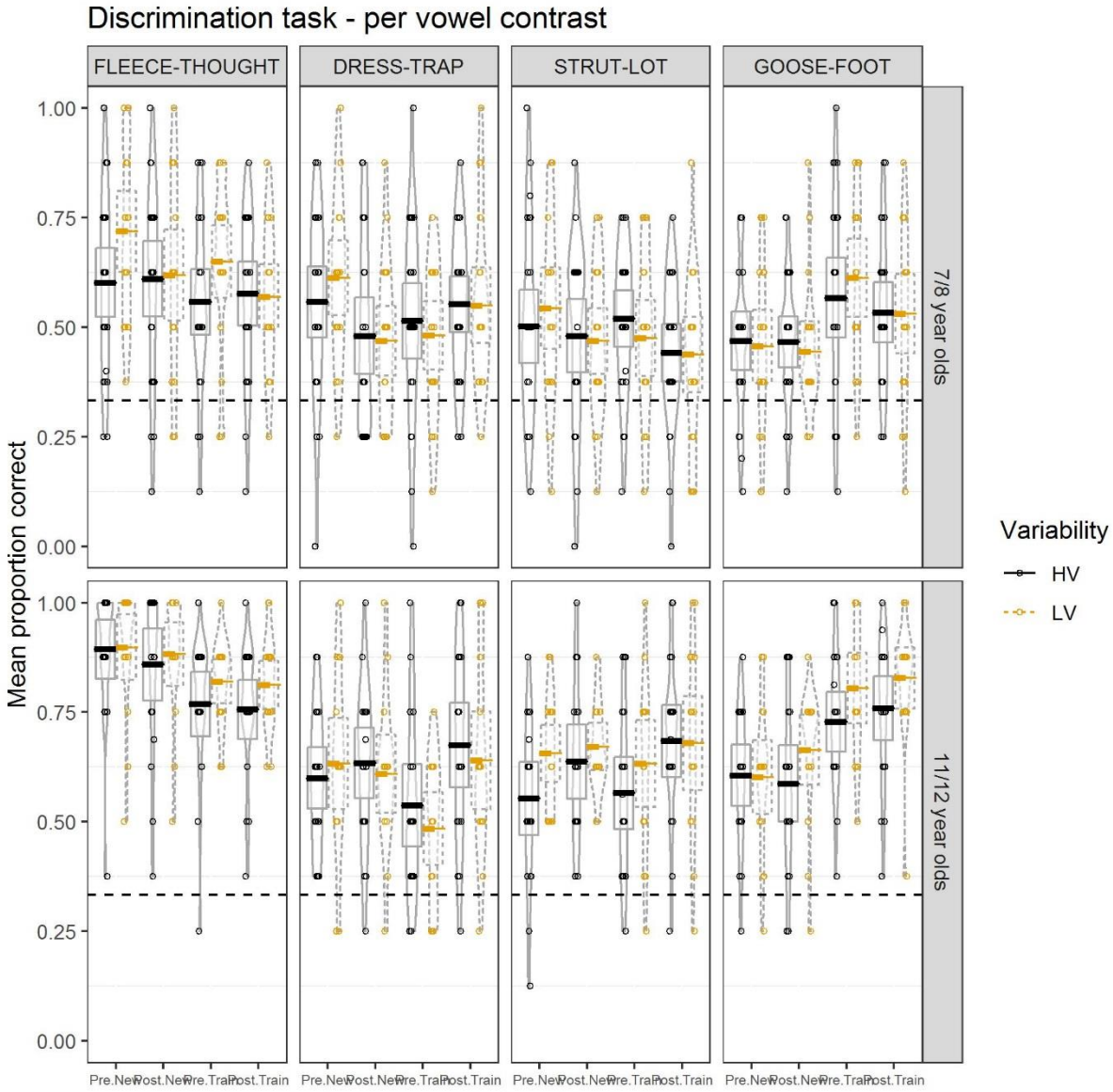
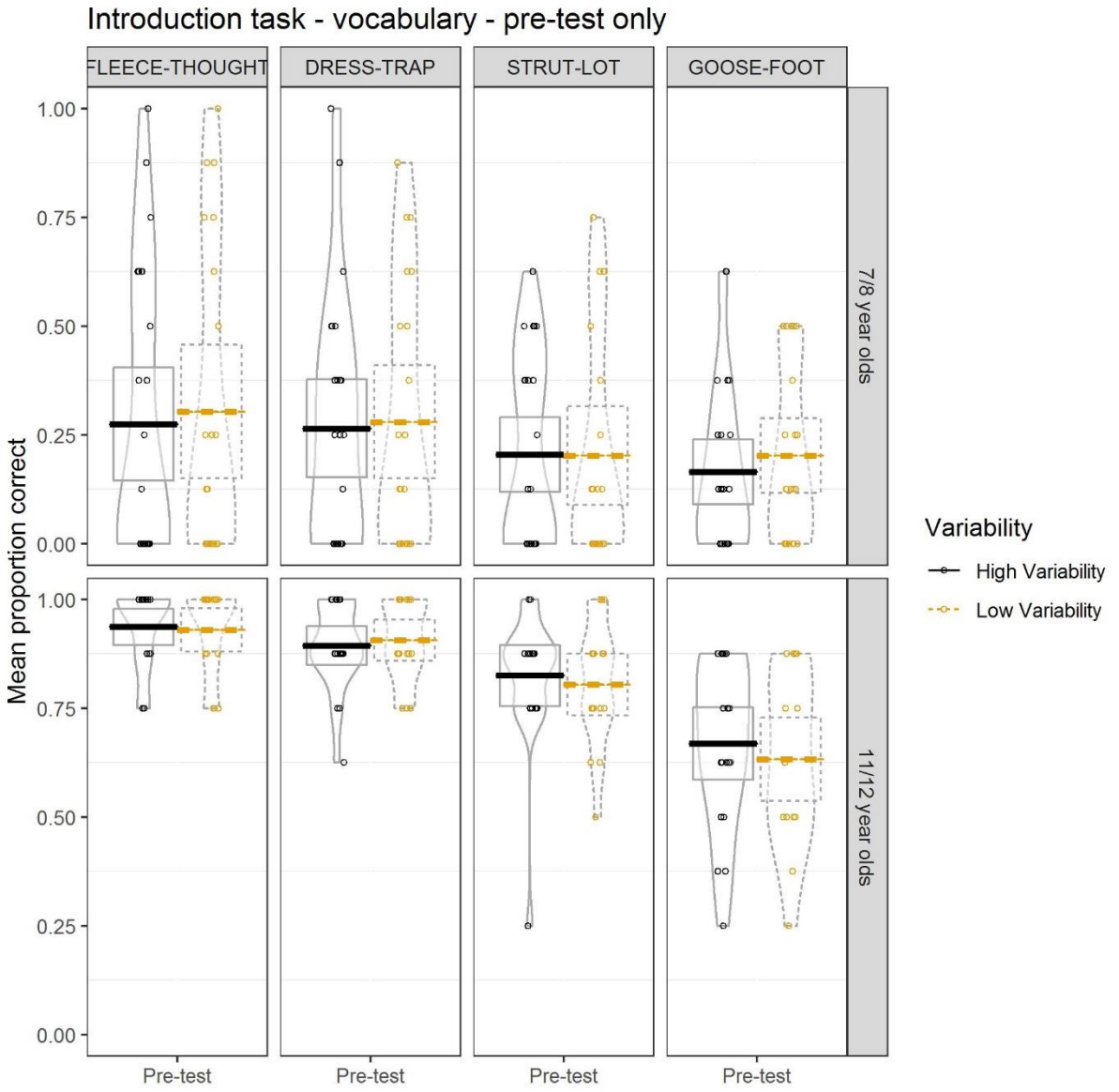
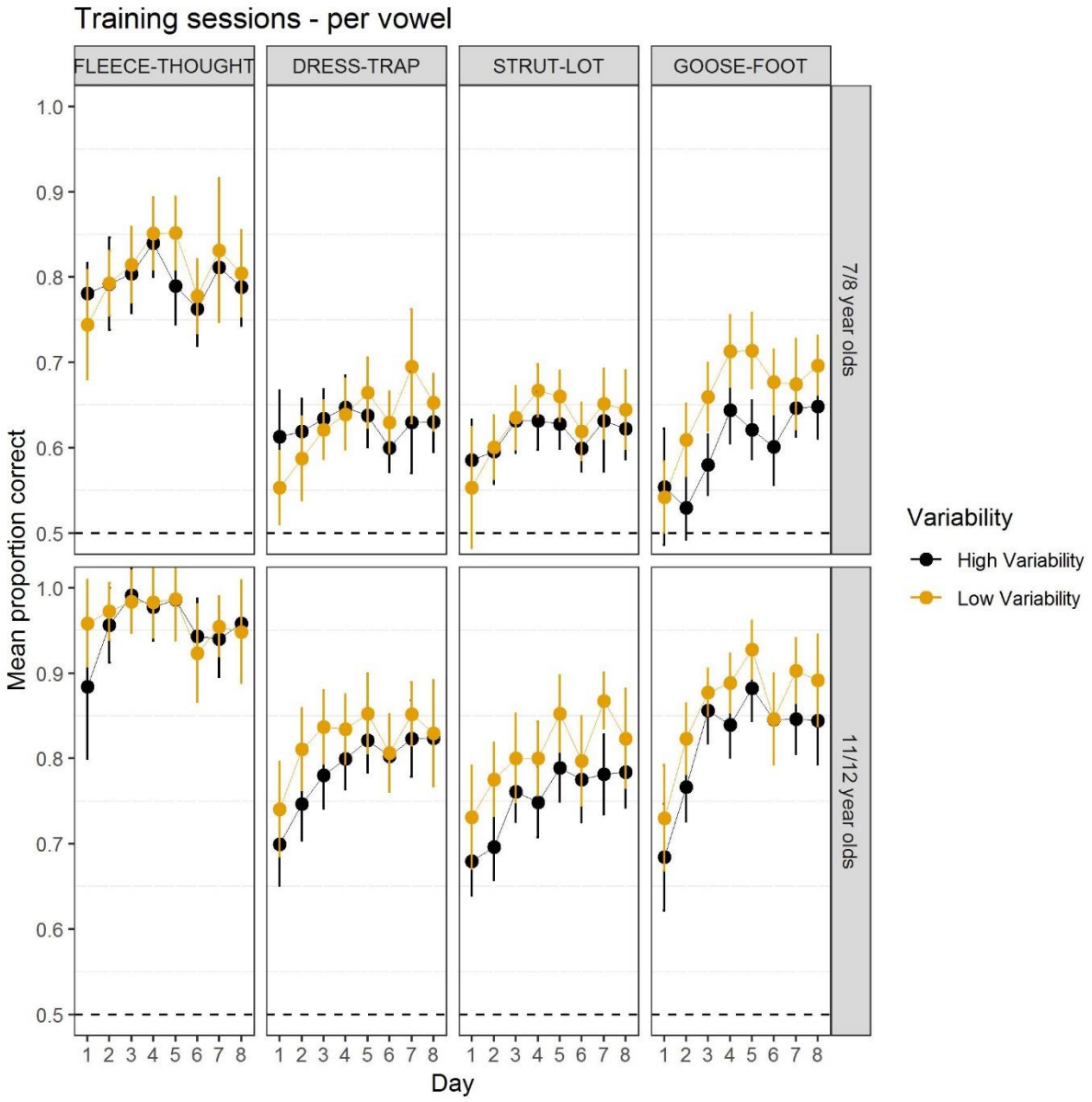


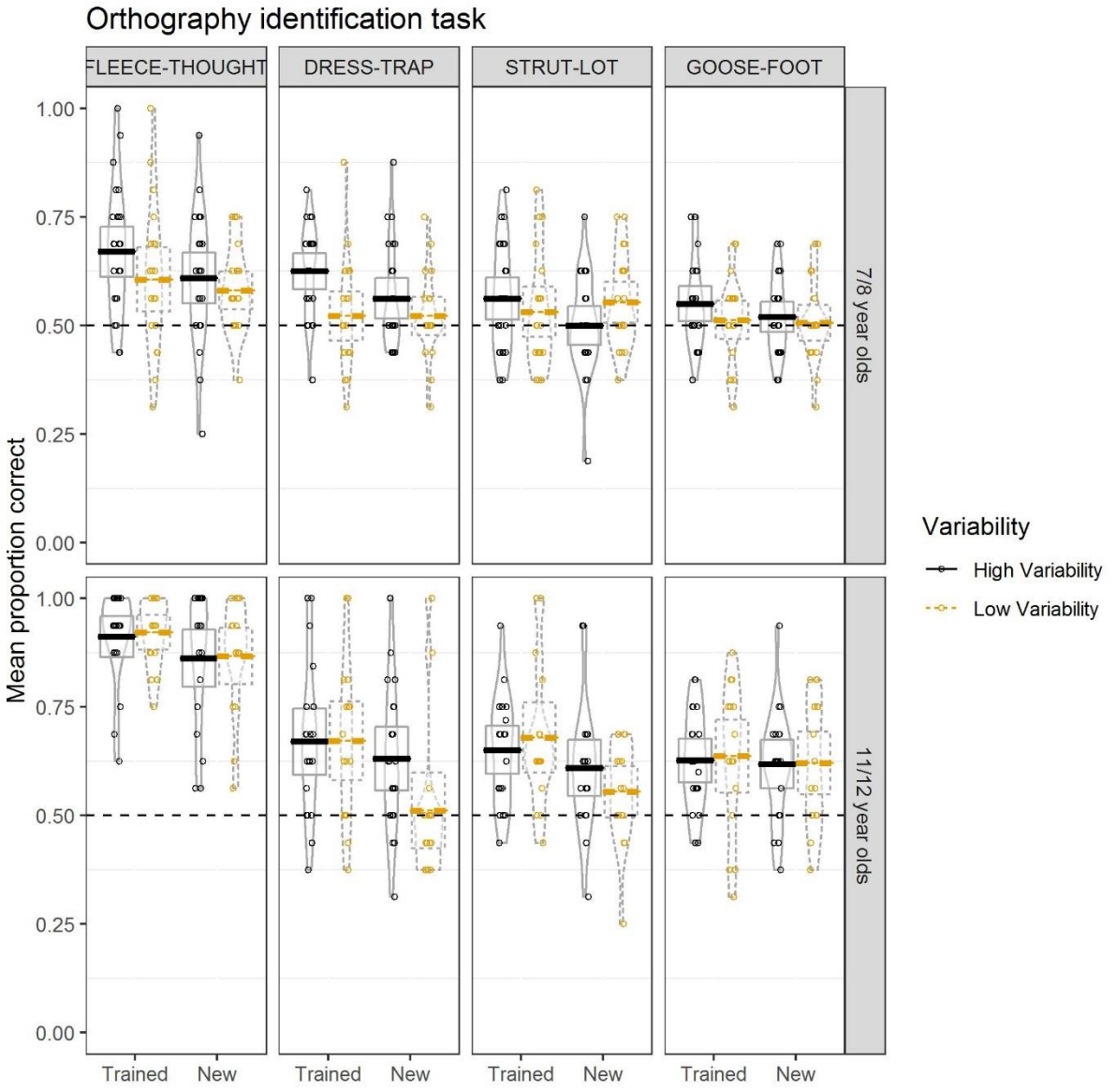
Figure 64. Pilot results from the Category Boundary task, investigating the category boundary for the PEN-PAN contrast.

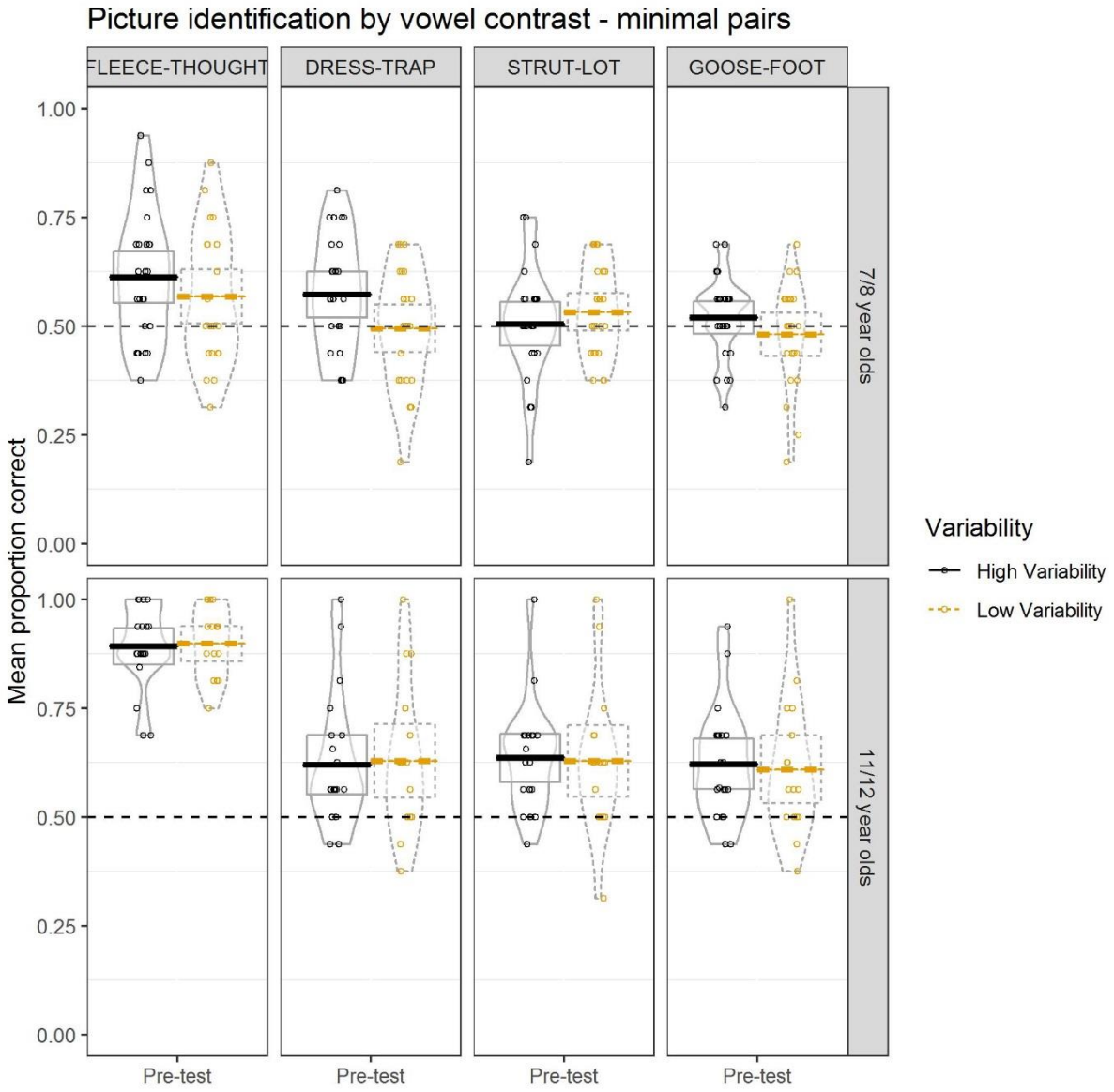
Appendix VIII. Study 2 - Task performance split by vowel contrast

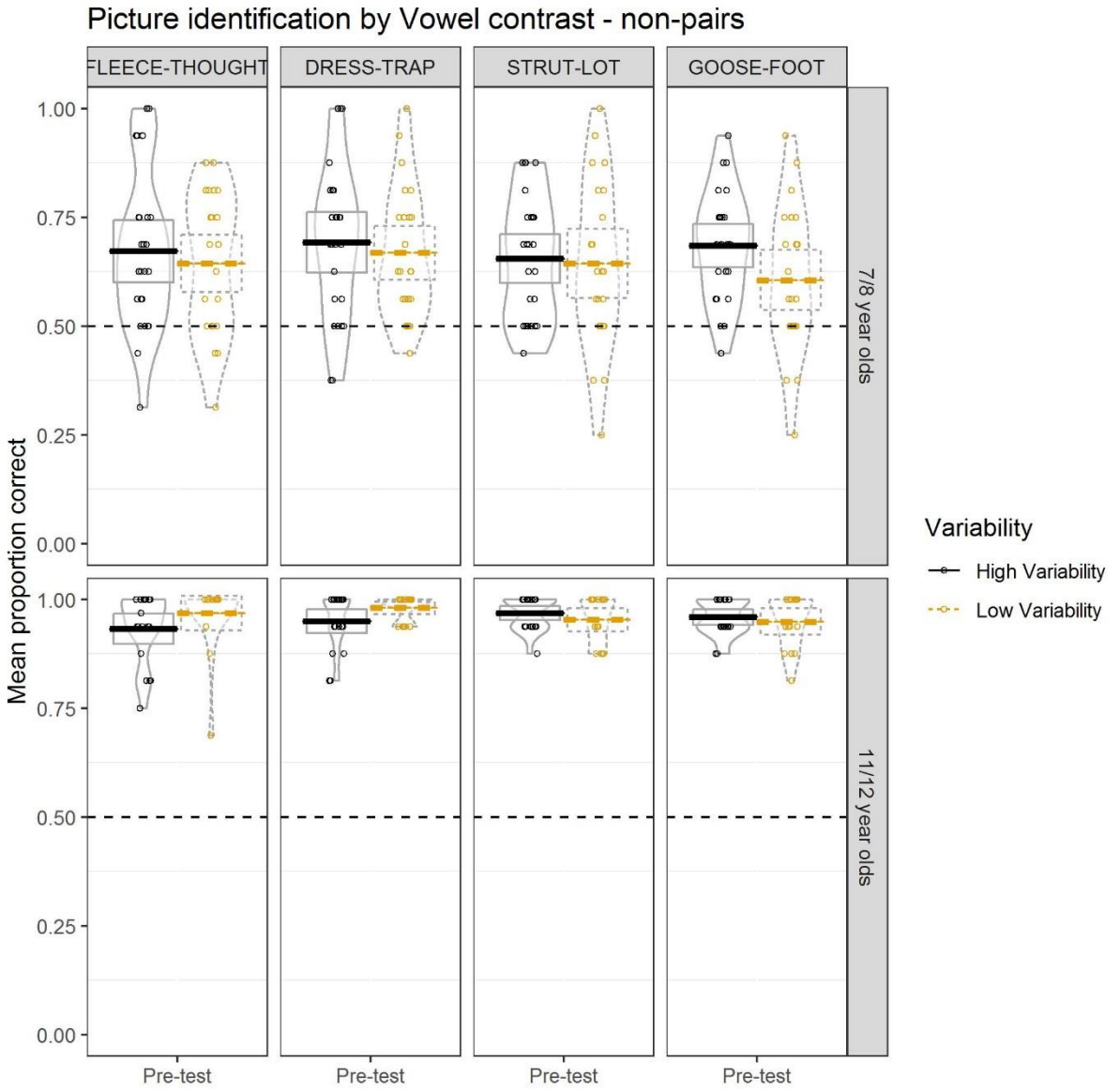



Introduction task - vocabulary - pre-test only
FLEECE-THOUGHT
DRESS-TRAP
STRUT-LOT
GOOSE-FOOT
7/8 year olds
11/12 year olds
Mean proportion correct
1.00
0.75
0.50
0.25
0.00
Pre-test
Variability
High Variability
Low Variability


Training sessions - per vowel
FLEECE-THOUGHT
DRESS-TRAP
STRUT-LOT
GOOSE-FOOT
7/8 year olds
11/12 year olds
Mean proportion correct
1.0
0.9
0.8
0.7
0.6
0.5
1 2 3 4 5 6 7 8
Day
Variability
High Variability
Low Variability


Orthography identification task
FLEECE-THOUGHT
DRESS-TRAP
STRUT-LOT
GOOSE-FOOT
7/8 year olds
11/12 year olds
Mean proportion correct
1.00
0.75
0.50
0.25
0.00
Trained
New
Variability
High Variability
Low Variability


Picture identification by vowel contrast - minimal pairs
FLEECE-THOUGHT
DRESS-TRAP
STRUT-LOT
GOOSE-FOOT
7/8 year olds
11/12 year olds
Mean proportion correct
1.00
0.75
0.50
0.25
0.00
Pre-test
Variability
High Variability
Low Variability


Picture identification by Vowel contrast - non-pairs
FLEECE-THOUGHT
DRESS-TRAP
STRUT-LOT
GOOSE-FOOT
7/8 year olds
11/12 year olds
Mean proportion correct
1.00
0.75
0.50
0.25
0.00
Pre-test
Variability
High Variability
Low Variability

Appendix IX. Study 2 – Production: Intra-rater reliability scores split out by vowel

Non-native speakers		Native speakers	
Vowel	Percentage agreement (vowel selection)	Vowel	Percentage agreement (vowel selection)
FOOT (37 items)	64.9%	**FOOT** (6 items)	83.3%
GOOSE (33 items)	63.6%	**GOOSE** (2 items)	100%
DRESS (31 items)	67.7%	**DRESS** (9 items)	100%
TRAP (24 items)	62.5%	**TRAP** (5 items)	100%
STRUT (36 item)	77.8%	**STRUT** (1 item)	100%
LOT (32 item)	59.4%	**LOT** (1 item)	100%
FLEECE (34 items)	79.4%	**FLEECE** (7 items)	100%
THOUGHT (22 items)	86.4%	**THOUGHT** (7 items)	85.7%
Overall (249 items)	69.9%	**Overall** (38 items)	94.7%

The vowel selection rating is the percentage agreement on the specific choice of vowels out of the twelve vowel options. This score is moderate to high in all trials produced by the non-native children, and high to perfect in the native speaker control trials. Seemingly, the control vowels FLEECE and THOUGHT are most consistent within the non-native trials (which is sensible given that they were easier to produce for the Dutch children).