Note: this is the authors final version, submitted 18/02/19

The role of input variability and learner age in second language vocabulary learning

Ruta Sinkeviciute[a], Helen Brown[b], Gwen Brekelmans[a], & Elizabeth Wonnacott[a]

[a] Psychology and Language Sciences, University College London, Gower Street, London, WC1E 6BT, UK

[b] Division of Psychology, Nottingham Trent University, Burton Street, Nottingham, NG1 4BU, UK

Correspondence concerning this article should be addressed to Elizabeth Wonnacott, Psychology and Language Sciences, University College London, Gower Street, London, WC1E 6BT, UK. Email: e.wonnacott@ucl.ac.uk

Abstract

Input variability is key in many aspects of linguistic learning, yet variability increases input complexity, which may cause difficulty in some learning contexts. The current work investigates this trade-off by comparing speaker variability effects on L2 vocabulary learning in different age-groups. Existing literature suggests that speaker variability benefits L2 vocabulary learning in adults, but this may not be the case for younger learners. In this study, native English-speaking adults, 7-8 year-olds, and 10-11 year-olds learned six novel Lithuanian words from a single speaker, and six from eight speakers. In line with previous research, adults showed better production of the multi-speaker items at test. No such benefit was found for either group of children either in production or comprehension. Children also had greater difficulties in processing multiple-speaker cues during training. We conclude that age-related capacity limitations may constrain the ability to utilise speaker variability when learning words in a new language.

Key words: Input Variability; Word Learning; Second Language Learning; Child Language Learning

Introduction

Any model of language learning must explain how learners cope with the variability which characterises human language at all levels of description. Much research has explored the idea that encountering variability aids learning by focusing the learner on the invariant, and thus linguistically important aspects of the input. For example, experimental and computational research suggests that token variability plays a role in driving syntactic and morphological generalisation in child language learning (e.g., Bybee, 1995; Gomez, 2002; Plunkett & Marchman, 1991; Wonnacott, Boyd, Thomson & Goldberg, 2012). There is also evidence that lower level acoustic variability (e.g., from tokens produced by varying speakers) plays a role in lexical learning, both in promoting the learning of lexically relevant phonetic contrasts (Lively, Logan & Pisoni, 1993; Logan, Lively & Pisoni, 1991; Rost & McMurray, 2009, 2010) and more generally (Barcroft & Sommers, 2005, 2014; Richtsmeier, Gerken, Goffman & Hogan, 2009; Sommers & Barcroft, 2007, 2011). The current work further explores the role of speaker variability (sometimes referred to as "talker variability") in the relatively understudied domain of child second language learning. Below, we review the literature concerning speaker variability in the area of child first language learning, followed by that on adult second language learning. Finally, we discuss a single study on speaker variability in the domain of child second language learning.

Speaker Variability in Child First Language Learning

One line of evidence suggesting a benefit of speaker variability comes from infants in the early stages of word learning (~14 months). A surprising finding is that even if infants have apparently mastered a particular phoneme contrast in their native language they may have difficulties learning novel words which differ only in this contrast. For example, Stager and Werker (1997) found that although 14-month-olds could discriminate /*b*/ and /*d*/, they could not successfully differentiate the novel minimal pair words /*bɪ*/ and /*dɪ*/ in a word learning context. This effect has been demonstrated many times (see Werker & Curtin, 2005, for a review). Critically, however, Rost and McMurray (2009) demonstrated that when the novel minimal pair words (/*buk*/ and /*puk*/ in their study) were spoken by multiple speakers, infants of the same age were successful in mapping each novel minimal-pair word onto a novel object. Further studies and computational modelling suggest that this difference is due to the fact that when the words are spoken by a single speaker, consistent cues from that speaker become associated with the object and this occurs at the expense of phonetically relevant cues (Apfelbaum & McMurray, 2011; Rost & McMurray, 2010; cf. Galle et al., 2015, for evidence that the benefit of variability does not rely on multiple speakers per se, since similar effects are seen from a single speaker who deliberately varies mean pitch, pitch contour, and duration of word tokens). Note that this account assumes that word learning is an associative process in which even linguistically irrelevant cues may be incorporated into lexical representations, at least in the early stages of learning.

There is also evidence that speaker variability may benefit word learning in older

children. Richtsmeier et al. (2009) taught 4-year old English-speaking children novel English nonce words (i.e., adhering to English phonology and phonotactics) associated with novel animal pictures. Words that had been presented in multiple voices were later repeated faster and more accurately than words that had been presented in a single voice, once again suggesting that variability may play a role in boosting lexical learning, even when there are no fine-grained phonetic distinctions to dissociate.

Speaker Variability in Adult Second Language Learning

The literature on adult second language (L2) learning also suggests a beneficial role of input variability (particularly speaker variability) in both phonetic and lexical learning.

In the domain of phonetic learning, early work by Strange and Dittman (1984) demonstrated that adults trained on a non-native phoneme contrast using synthesised tokens from a single phonetic environment were not able to generalise learning to untrained words or speakers. However, seminal work by Logan et al. (1991) demonstrated that when adults were trained using varied minimal pair stimuli (i.e., where the contrast occurred in varied syllabic environments) spoken by multiple speakers they showed post-test improvements for both untrained words and speakers. Lively et al. (1993) replicated this finding; however, in a follow-up experiment, adults trained using varied minimal pair stimuli spoken by a single speaker (i.e., there was item but not speaker variability) showed post-test improvements for the trained speaker, but not for an untrained speaker, suggesting a specific role for speaker variability in high-variability training. These studies indicate that listeners encode indexical information in a

way that facilitates the later distinction between challenging L2 phonemic contrasts. A later study by Bradlow, Akahane-Yamada, Pisoni and Tohkura (1999) found that comprehension and production benefits from high-variability training were maintained three months post-training. High variability phonetic training has since become a standard methodology in the field, and has been applied to other contrasts (e.g., Iverson, Pinet & Evans, 2012).

It is worth noting, however, that the original study by Logan et al. (1991) had a very small sample size (the generalisation tests were administered to only three of the participants; see also Pruitt, 1993) and only a few studies have returned to directly comparing multi-speaker and single-speaker input for phonetic training. One such study was conducted by Clopper and Pisoni (2004) who found a benefit for high- over low-variability training in the related area of dialect categorisation. Another study which did look at phonetic contrasts is Sadakata and McQueen (2013): learners trained with less variable input, comprising many repetitions of a limited set of words recorded by a single speaker, showed less generalisation than those trained with fewer repetitions of a more variable set of words recorded by multiple speakers. Together these studies suggest a benefit of speaker variability in the domain of phonetic learning.

Turning now to lexical learning, Barcroft and Sommers (2005) investigated the role of two sources of acoustic variability: speaker variability and speaking style. In one experiment, English-speaking adults learned 24 Spanish words where eight words were learned in each of the three conditions: no variability (each word was produced six times by one speaker), moderate variability (each word was produced twice by each of three speakers) and high variability (each word was produced once by each of six speakers). Learning was assessed using a production

(picture-to-Spanish word) test and a comprehension (Spanish word-to-English word) test. Reaction times and accuracy scores for both tests showed that L2 vocabulary learning improved systematically, moving from low to moderate and from moderate to high variability conditions. The same pattern of results was found when the speaker was held constant and variability in speaking style was similarly manipulated. Sommers and Barcroft (2007, see also Barcroft and Sommers, 2014) further established that manipulating variability in speech rate had similar benefits to speaker and speaker style, whilst manipulating amplitude and fundamental frequency (F0) did not; however speakers of a tonal language (where contrasts in F0 are lexically relevant) *did* show a variability effect for F0. The authors suggest that variability affects learning for those cues that are relevant to word recognition in the learners' first language (L1).

Further experiments by Sommers and Barcroft (2011) provided evidence against an explanation in which the benefit of high variability arises from the greater cognitive effort imposed by high variability input, due to the more difficult encoding demands (the "cognitive effort" hypothesis). This hypothesis predicts that L2 vocabulary learning should be superior under any manipulations which increase encoding difficulty. However, Sommers and Barcroft found that this was not the case for poor signal to noise ratios; instead, Barcroft and Sommers (2005) explain the benefit of acoustic variability in their data in terms of an exemplar-based framework whereby indexical information from all encountered examples may be retained in the early stages of learning. Thus, when words are encountered from multiple speakers/voice types, learners incorporate a wider variety of cues into their representations, allowing them to form more "associative hooks", resulting in more robust representations for the novel words. They

refer to this hypothesis as the elaborative processing hypothesis, contrasting it with a view in which variation is normalized and does not affect processing (Barcroft, 2001).

To summarise, there is evidence from both child L1 acquisition and adult L2 acquisition, that speaker variability benefits lexical learning. However, these phenomena have been interpreted somewhat differently in the different literatures: researchers exploring adult L2 acquisition (Barcroft & Sommers, 2005) have suggested that speaker-specific cues make novel lexical representations more robust; developmental researchers (Apfelbaum & McMurray, 2011; Rost & McMurray, 2009, 2010) have argued that varying speaker cues may prevent consistent speaker cues becoming associated with particular objects at the cost of phonetically relevant cues. An alternative account, not directly discussed by either set of authors, is provided by discriminative learning models in which learning is a process by which prediction error is used to discriminate uninformative cues and to reinforce informative cues. In this view, the presence of varying, non-predictive speaker cues may assist in decontextualizing lexical representations – which in this view are essentially optimal predictive codes (Ramscar & Baayen, 2013; Ramscar, Yarlett, Dye, Denny & Thorpe, 2010). Regardless of which of these specific theoretical approaches is correct, it seems likely that there is a single explanation for the benefit of speaker variability in L1 and L2 lexical learning, and that this phenomena can occur across different age groups. From this perspective, we would predict that a similar benefit should be found in child L2 learning.

Speaker Variability in Child Second Language Learning

Despite the fact that child L2 learning is highly common (due to immigration, bilingual communities and schooling, e.g., Eurostat, 2015) the literature in this area is notably sparse compared with that on child L1 and adult L2 acquisition, and is largely focused on differences between early and late L2 learners. Aside from this interest in the role of age of acquisition, relatively little research explores which factors promote better L2 learning in children, and whether these are the same as for adults. Only one study (to our knowledge) has addressed the question of whether high speaker variability training benefits L2 learning in children. Giannakopoulou, Brown, Clayards and Wonnacott (2017) trained Greek-speaking 8-year-olds and adults on the non-native English /i/-/ɪ/ contrast in one of two conditions – low (one speaker input) or high (four speaker input, with speaker changing on a trial-by-trial basis) speaker variability. All participants were current learners of English and (due to the populations available for sample) adult participants' starting level of proficiency was advanced whilst children's was basic. Participants learned the contrasts using a two-alternative forced choice task (2AFC, e.g., hear *ship*, choose between pictures of a *sheep* and a *ship*, where the foil picture was always the minimal pair item). During training (10 sessions), accuracy was higher in the low speaker variability condition for both age-groups, likely due to the fact that hearing the same speaker in each trial did not require constant adaptation to a different speaker. However, the critical test was whether high or low variability training would be most beneficial in generalisation to new words or speakers. A three-interval oddity discrimination test (administered both pre- and post-training) showed that discrimination improved in both adults and children in both variability conditions.

However, there was no benefit of high over low variability training for either age-group. Instead, for adults, the extent of improvement from pre- to post-test between the two variability conditions was not statistically significant, whilst children actually showed *more* improvement in the low variability condition, for both trained and untrained words. However, adults were close to ceiling in the discrimination task, potentially masking a high variability benefit (which was numerically present). For children, the apparent low variability advantage could have been due to accidental differences at pre-test (the low variability group started off with lower performance, giving more room for improvement). Still, there was clearly no evidence of a high variability benefit for children in this study.

Although Giannakopoulou et al. (2017) were focused on the question of phonetic learning, the contrast was embedded in a word learning task (matching L2 words to pictures) and lexical knowledge was tested pre- and post-training using a four-alternative forced choice task in which participants saw a picture and selected which of four English words went with it. Adults were at ceiling at this test; children were not and showed pre-to post-test improvement. However, they again showed no benefit of high variability, in contrast to the results from the adult L2 word learning studies described above.

In sum, the results of Giannakopoulou et al. (2017) found no evidence that 8-year-olds benefited from high speaker variability input for training either novel phonetic discrimination or lexical learning. This is in contrast to the studies reviewed above for both discrimination and vocabulary training in adult learners, and also for children in a L1 learning context. One possibility discussed by Giannakopoulou et al. is that the variety of indexical cues present in the

multiple speaker condition may have placed an increased burden on processing. This might be particularly difficult for children, since they have lesser phonological working memory capacity than adults. This increased burden may have outweighed any benefit to be gained from variability. However, the ceiling effects observed in adult learners make these results difficult to interpret. In order to establish that children do indeed differ from adults in their ability to benefit from high variability input, it is therefore necessary to compare adult and child learners with a more comparable starting point, and to explore whether children show a variability advantage in a context where it is seen with adults.

The Current Study

The current study focuses on the domain of L2 vocabulary learning and aimed to answer the question of whether high speaker variability benefitted L2 vocabulary learning in both children and adults. The paradigm is loosely based on the methodology established by Barcroft and Sommers (2005), and had three phases – training, production testing and comprehension testing – though several modifications were made.

First, a different modern language was used: Lithuanian. Second, to mitigate against possible floor effects in children's responses participants learned only 12 words (instead of 24), with each word repeated eight times (instead of six). Additionally, we used only two (not three) experimental conditions – high versus low variability. Third, we used a computerized 2AFC training task (as in Giannakopoulou et al., 2017), rather than the more passive 'look and listen' task used by Barcroft and Sommers (2005). This allowed us to collect data from participants

during training and thus see whether they were able to identify the pictures the first time, how their performance was affected by the variability in the input and how this differed across learners, which may be important in shedding light on how the different age-groups later fared at test. Additionally – in anticipation of the production test – we asked participants to repeat the words during training. Finally, the comprehension test was altered to suit child participants better: rather than L2-to-L1 translation, L2 word-to-picture recall was tested, where participants were asked to identify the correct picture given the complete set of items tested in the experiment.

We tested two age-groups of children in the current work: 7-8 year-olds and 10-11 year-olds. The rationale for testing 7-8 year-olds is that this is the age at which children now begin foreign language learning in UK schools, making the results of direct interest to potential UK educators. It is also close to the age of the participants in Giannakopoulou et al. (2017). Adult participants were included for comparison purposes. Older children were included in case younger children differed from adults, to begin exploring at what age children pattern like adults (10-11 year olds were targeted since they could be recruited from the same UK primary schools).

Since adults usually outperform children in word learning tasks (e.g., Henderson, Weighall, Brown & Gaskell, 2013), we predicted that adults would be more accurate than either age-group of children both in training and at test. We also predicted that older children would outperform younger children in each test. Most important are our predictions regarding the variability manipulation. These differed for data collected in the training and the testing phases. For training, we predicted there would be higher accuracy with low variability items (a low

variability benefit) for all age-groups. This is predicted since participants only have to attune to a single speaker for these words, easing their recognition throughout the training task. The prediction is also in line with the findings of Giannakopoulou et al. (2017) who used a similar training task. In contrast, at test, participants are asked to generalise away from the trained talker/talkers, either producing the words themselves (Production test) or understanding them with a new talker (Comprehension test). Thus, based on Barcroft and Sommers (2005), we predicted that adults would recall the high variability words significantly better than the low variability words due to the more robust representations formed for the words learnt from multiple speakers. For the two child age-groups, our predictions at test are less clear based on the previous literature; however, we tested the hypothesis that they will show the same high variability benefit effect as adults.

Method

Participants

Thirty-two 7-8 year-olds (*M*=7;10, *SD*=4.2 months, 22 female), thirty-two 10-11 year-olds (*M*=11;0, *SD*=5.6 months, 19 female), and 32 adults (*M*=20;6, *SD*=3.4 years, 27 female), participated in the study.[1] Participants were native English speakers. However, seventeen 7-8

[1] Since our sample is unbalanced with respect to gender, we conducted additional analyses (see https://osf.io/d2gkh/?view_only=dc90adf4ab724d00807536767614abb8) to see if including gender in the models would change the results. There was no consistent pattern with respect to gender, and key results reported in the text

year-olds, ten 10-11 year-olds and 18 adults were simultaneous bilinguals. None of the additional languages spoken by the participants were related to the target language used (Lithuanian, see Appendix A). The children were all learning French as a foreign language at school.[2]

For children, parents/guardians provided informed consent prior to the session, and children gave verbal consent to participate. Children received a certificate and stickers for their participation. Adult participants were recruited through the SONA participant pool at University College London. Informed consent was obtained at the beginning of the session. Twenty-four adults received course credit and the remainder were paid for their participation.

None of the participants had any language or hearing impairments. The participants were not familiar with any Lithuanian words prior to the study and were blind to the aims of the study.

Stimuli

The stimuli consisted of 12 Lithuanian nouns which were singular, countable, and unambiguous in both Lithuanian and English. The age of acquisition of the selected English nouns was between 3 and 4 years of age (Kuperman et al., 2012). Three of the words were

---

remained un-changed.

[2] We conducted analyses (see https://osf.io/d2gkh/?view_only=dc90adf4ab724d00807536767614abb8) to see if including bilingualism as a fixed effect in any of the three statistical models reported in the results section (i.e., for training, production and comprehension) added significant variance to the model: it did not in any case. In addition, inspection of the coefficients confirmed that the pattern of significances for experimental variables remained the same.

English cognates/near cognates (*klounas* clown; *tigras* tiger; *tortas* cake [tart]). Counterbalancing was used such that these (and all other items) occurred equally as high/low variability across participants[3]. Specifically, the words were divided into two lists (Table 1) with word frequency roughly balanced according to English translation word frequency: mean word frequency 604 words/million for List 1, and 523 words/million for List 2 (Children's Printed Word Database, 2002).

[3] Since we have bilingual participants, we also checked if any other items are cognates in any of their languages. The only cognate items were (i) the same three as are cognates in English (which are cognates in many of the other languages – *tortas*: Spanish, Swedish, French, Polish, Italian, Portuguese, Hungarian, Lithuanian, German, Albanian, Greek, Russian, Punjabi; *klounas*: Spanish, Swedish, Italian, German, Filipino, Albanian, Lithuanian, Polish, Greek, Russian, Malayalam; *tigras*: Swedish, German, Albanian, Lithuanian, Spanish, Italian, Filipino, French, Portuguese, Hungarian, Polish, Greek, Russian) (ii) *karve*, which is a cognate in Russian (*корова*) and Polish (*krowa*) (iii) *knyga* which is a cognate in Russian (*книга*) (iv) *meska*, which is a cognate in Polish (*mis*). For the words which are not also cognates in English, these items could potentially bias our analyses. We therefore repeated the analyses for Production (where we see the clearest age differences) with these items removed for the relevant participants (i.e., one adult Russian and Polish speaker, three 7-year-old Polish speakers). The key pattern of results was unchanged. Analyses are included in https://osf.io/d2gkh/?view_only=dc90adf4ab724d00807536767614abb8.

Table 1: *Word Lists*

List 1		**List 2**	
Lithuanian	***English***	***Lithuanian***	***English***
klounas /'klɔʊnɐs/	clown	tigras /'tɪgrɐs/	tiger
meška /mɛʃ'kɐ/	bear	višta /vɪʃ'tɐ/	chicken
tortas /'tɔ:rtɐs/	cake	voras /'vɔ:rɐs/	spider
medis /'mɛ:dɪs/	tree	namas /'na:mɐs/	house
kiškis /'kɪʃkɪs/	rabbit	karvė /'ka:rve:/	cow
raktas /'ra:ktɐs/	key	knyga /kni:'gɐ/	book

Five male and five female native Lithuanian speakers aged 20-25, who had lived in Lithuania at least until the age of 16, recorded the words using a normal intonation in a sound-attenuated room using a sampling rate of 44.1 kHz. Peak amplitude of each sound file was normalized using Praat (Boersma & Weenink, 2017). Pilot testing demonstrated that a native Lithuanian speaker (not involved in the study prior to pilot testing) successfully identified all 120 productions. Clipart cartoon pictures representing the 12 words were selected (one per picture) from free online clipart databases. Audio and picture stimuli are available on OSF (see https://osf.io/d2gkh/?view_only=dc90adf4ab724d00807536767614abb8).

Design

The experiment used a within-participants design. The independent variable was input variability (high vs. low, with one word list assigned to each level of input variability) and the dependent variables were production and comprehension test accuracy scores. During training, words in the list assigned to the high variability condition were exemplified by eight speakers (four male, four female) while words in the list assigned to the low variability condition were exemplified by only one of those speakers. Each participant was assigned to one of 16 versions of the same training task that counterbalanced the word lists as well as which speaker exemplified the low variability word list, thus controlling both for potential differences in word difficulty across the lists, and for the intelligibility of different speakers assigned to be the single LV speaker.

Procedure

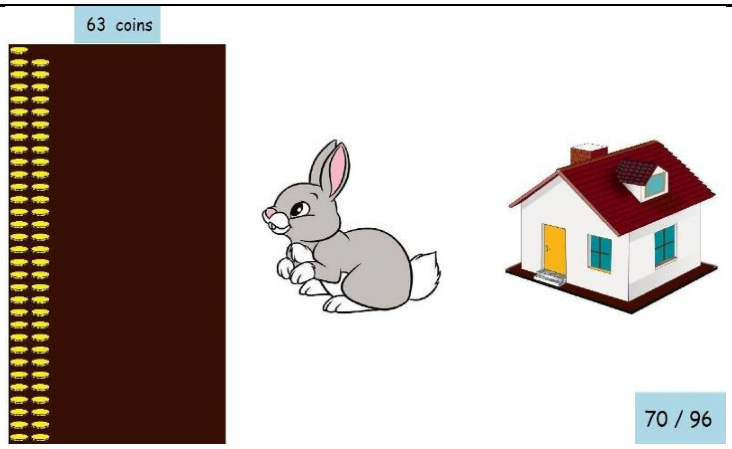
Participants were tested individually using a Samsung laptop, and Sennheiser HD 201 headphones in a sound attenuated room (adults) or a quiet area of their school (children). The experimental tasks were administered using ExBuilder software (a custom-built software package developed at the University of Rochester). The experiment consisted of three parts administered in a fixed order: training, production test, and comprehension test (without breaks). The experiment lasted approximately 30 minutes.

Training Task

Participants were told that they were going to play a language game, aiming to collect as

many coins as possible. In each trial participants saw two pictures and heard a Lithuanian word (Figure 1). Participants repeated the word aloud and then clicked on the picture that they thought corresponded to the word. Regardless of response accuracy, the incorrect picture then disappeared and the word was repeated by the same speaker. Each time the correct picture was selected the participant received a coin. Each word was the target word eight times resulting in 96 trials in total. The foil for any target was randomly selected from both lists. The presentation order of the stimuli as well as the display position of the correct picture in each trial was randomised for each participant.

Training trial



Incorrect guess Correct guess

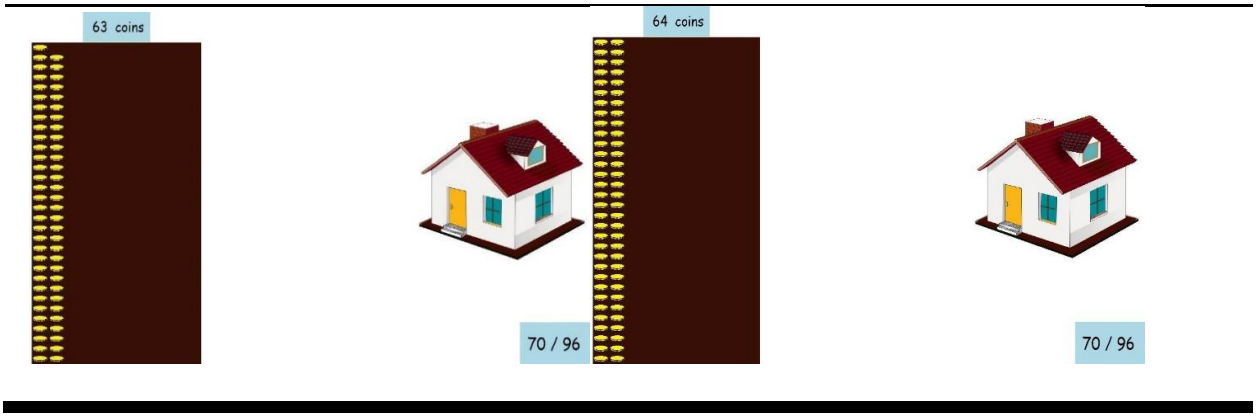


Figure 1: *An example of a training trial where the target word is 'namas' ('house').*

Production Test

Picture-to-word recall was tested first to avoid any additional exposure to the spoken Lithuanian words and to ensure that words in the low variability condition were heard only from one speaker prior to testing. Each picture was presented twice in a random order. Participants were asked to say the corresponding Lithuanian word for each picture. No feedback was given. Participants' responses were recorded and were later transcribed by a native Lithuanian speaker (the first author). At the time of transcription the coder was aware of the correct response word, but was blind to condition. A computer readable phonemic script (adapted from SAMPA, Wells, 1997) was used in order to be able to automatically compare productions with the correct response.

Comprehension Test

In the word-to-picture comprehension test each Lithuanian word was tested twice – once

using a novel female voice and once using a novel male voice. On each trial participants heard a Lithuanian word and selected the corresponding picture from a grid containing all 12 trained items (Figure 2). The presentation order of test items was randomised and no feedback was provided. For each individual participant, the pictures retained fixed positions within the grid throughout the comprehension task. Grid position was randomly determined for each participant.

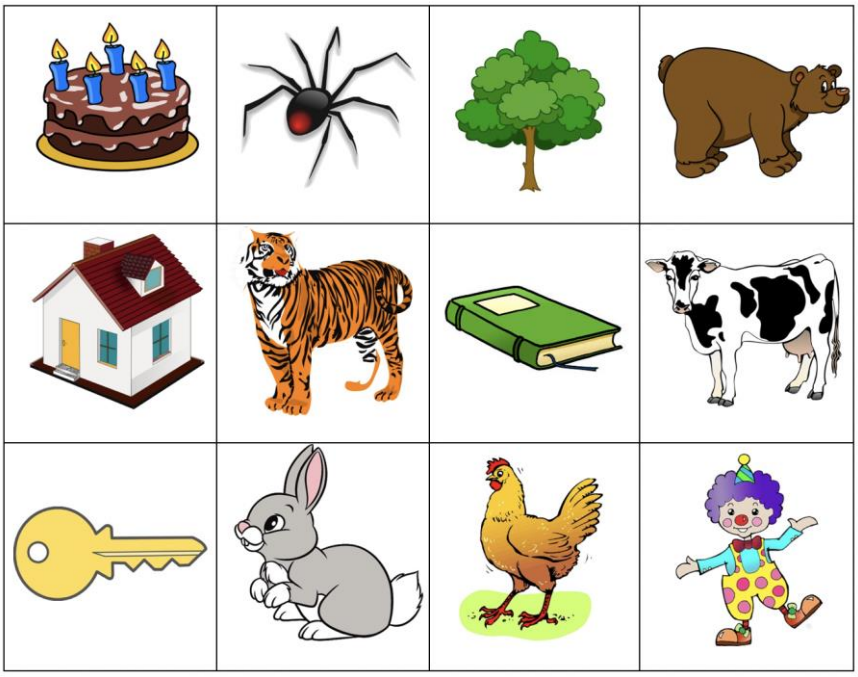


Figure 2: *An example of a grid during the comprehension test. The locations of the pictures were randomised for each participant. The same grid stays on the screen throughout the test. For each trial, participants heard a word (e.g., "namas" – "house") and clicked on the picture which they thought corresponded to the word.*

Results and Discussion

Overview

Data from each task (training, production, comprehension) were analysed separately and are reported below. For each task we conduct both frequentist and Bayes factor analyses. For the frequentist statistics, data were analysed in logistic mixed effects (LME) models predicting response accuracy (allowing binary data to be analysed with logistic models rather than as proportions (Jaeger, 2008; see also Baayen, Davidson & Bates, 2008; Quené & van den Bergh, 2008) using the lme4 package (Bates, Maechler, & Bolker, 2013) for R (R Development Core Team, 2010). Full details of each model are described below.

Additional Bayes Factor (BF) analyses were included for cases where we wished to explore whether there was support for the null hypothesis. A non-significant result ($p>.05$) in frequentist analyses is ambiguous as to whether there is actual support for the null over H1, versus insufficient evidence to distinguish H1 and the null (note this is the case even if the means are reversed; see Dienes, 2014, for further discussion of how null results are routinely misinterpreted in the literature). In contrast, *BFs* provide a measure of how strongly the data support the H1 over the null – and vice versa. In this experiment, we used *BF*s in cases where the predicted difference between variability conditions (i.e., Training: low > high; Test: high > low) was not found for a particular age-group. To compute the *BFs*, we follow the method advocated by Dienes (2008, 2015), and, given a data summary (a mean difference and *SE*) compare the likelihood of the null hypothesis H0 (no difference) and of the H1, where H1 is modelled as the

half normal with the *SD* set to an estimate of the predicted mean difference. Since there are no appropriate prior data (i.e., using sufficiently similar materials) on which to base the estimate of the predicted mean difference, instead we inform H1 using measures obtained from within the current experiment. For example, in the Production test, since we get the predicted effect of high > low variability with adults, when computing *BFs* for each child group, our estimate of the predicted mean difference is set to be the *adult* mean difference between conditions. To meet the assumptions of normality, we continue to work in log odds space and thus our mean differences and *SEs* are taken from coefficients in the corresponding LME models. *BFs* > 3 indicate substantial evidence for H1, and *BFs* <1/3 indicate substantial evidence for H0.

Training

Accuracy (whether the correct picture was selected or not) was recorded by the software. Figure 3 shows the proportion of correct responses in the training task, split by *age-group* and *variability-condition*.

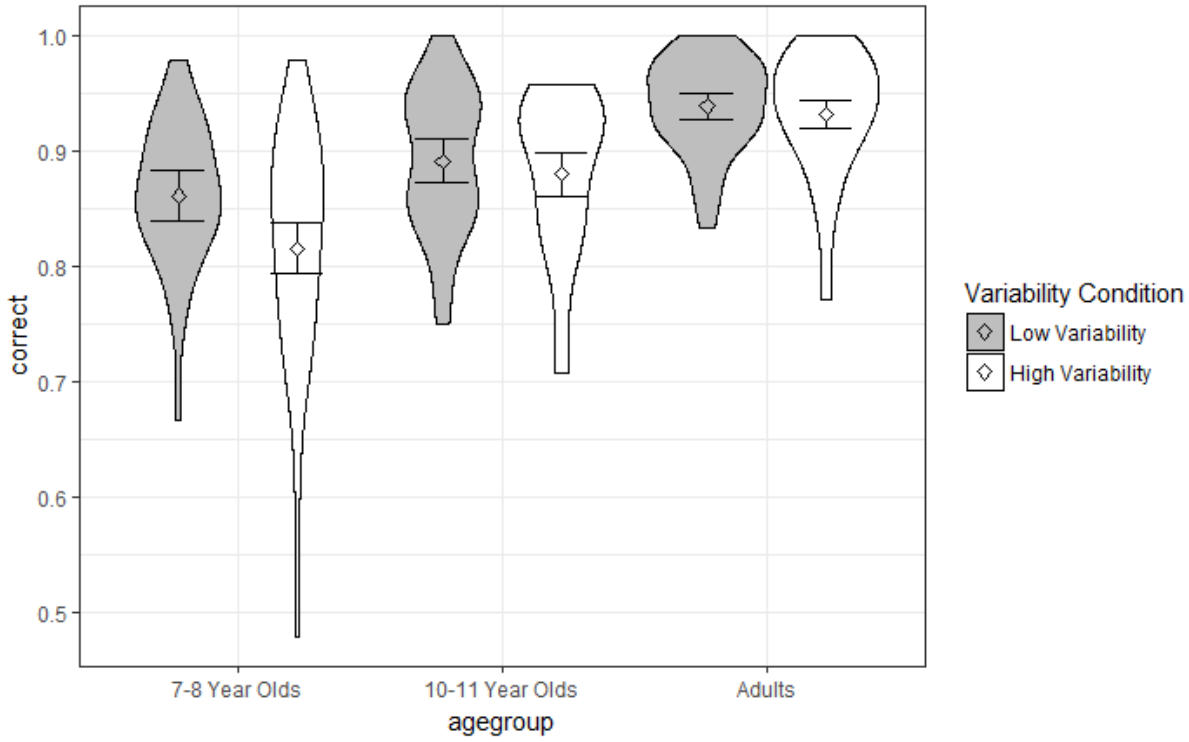


Figure 3: *Violin plot showing the proportion of correct response in the training task for high variability and low variability items. Shape shows the kernel probability density of participants' mean scores. Mean values are shown; error bars indicate 95% confidence intervals around the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Morey, 2008).*

Structure of Glmer Model

Predicted outcome: Response accuracy

Fixed factors*: Variability-condition* (low, high), *age-group* (7-year-olds, 10-year-olds, adults) and *trial-number* (numerical predictor 1 to 96), as well as all of the interactions between

them (regardless of whether they contributed significantly to the model – that is, we did not use stepwise model comparison). For both the discrete factor *variability-condition* and the numeric factor *trial-number,* we used a centered coding to reduce the effects of collinearity between main effects and interactions, and so that main effects were evaluated as average effects over all levels of the other predictors (rather than at a specified reference level for each factor). The factor *age-group* was coded using centered contrasts (again ensuring other effects were evaluated as averaged over all three levels of *age-group*). We ran the model twice with a different baseline each time, allowing us to inspect the contrasts between 7-year-olds versus adults, 10-year-olds versus adults, and 7-year-olds versus 10-year-olds.

Random effects: We included random intercepts for *participant* (96 levels), *word* (12 levels), and *speaker* (8 levels). For *participant,* we automatically included both a slope for *variability-condition* (the only within-participant factor) as well as the correlation between the slope and the intercept, that is, a full random slope structure for participants (Barr, Levy, Scheepers & Tily, 2013). A model which also had full random slopes structure for *word* and *speaker* (i.e., for each, slopes for *variability-condition, age-group* and the interaction between them, and correlations between slopes) did not converge. For these factors, we first identified the most complex model that would converge (by first removing correlations between slopes, then the interaction between the main effects). Further, to avoid an overly complex model, which has implications for power (Matuschek et al., 2017), we examined whether the slope structure could be reduced further by using a backwards model selection process whereby we started with the most complex model which converged, and reduced the model complexity until a further

reduction would imply a significant loss in the goodness-of-fit. Following Matuschek et al. (2017) the significance level of this model-selection criterion was specified as 0.2.

The final model had both by-word and by-talker slopes for *variability-condition* and *age-group*, with no correlations between slopes in either case. The final model and process by which slopes were reduced can be seen in the R analysis script available on OSF (see https://osf.io/d2gkh/?view_only=dc90adf4ab724d00807536767614abb8).

Results in Final Model

As expected, adults significantly out-performed both older and younger children, and older children outperformed younger children (7-year-olds, M=84%, SD=8%; 10-year-olds, M=89%, SD=6%; adults, M=94%, SD=4%; 10-year-olds vs. adults: β=1.14, SE=0.21, z=5.42, p<.001; 7-year-olds vs. adults: β=1.62, SE=0.24, z=6.77, p<.001; 10-year-olds vs. 7-year-olds: β=0.48, SE=0.2, z=2.39, p=0.017). There was a significant main effect of *trial-number* (β=0.86, SE=0.05, z=19.01, p<.001), reflecting participants increasing performance throughout training. Significant interactions between *age-group* and *trial-number* indicated that adults learned faster than both older children (β=0.6, SE=0.12, z=4.9, p<.001) and younger children (β=0.65, SE=0.12, z=5.45, p<.001), but older children did not learn significantly faster than younger children (β=0.05, SE=0.09, z=0.53, p=0.596).

There was no overall effect of variability (β=-0.1, SE=0.14, z=-0.72, p=0.472), however there was a significant interaction between variability and the contrast between 7-year-olds and adults (β=0.54, SE=0.27, z=2.02, p=0.043), though not between 10-year-olds and adults (β=0.25, SE=0.27, z=0.92, p=0.359) or 7-year-olds and 10-year-olds (β=0.28, SE=0.18, z=1.59, p=0.112).

We broke down the interaction between *age-group* and *variability-condition* by fitting a separate slope for variability for each age-group: 7-year-olds performed significantly better with low (*M*=86%, *SD*=6%) than high (*M*=82%, *SD*=11%) variability (β=-0.38, *SE*=0.16, *z*=-2.43, *p*=0.015); 10-year-olds showed no significant difference (low variability, *M*=88%, SD=6%; high variability, *M*=89%, *SD*=7%; β=-0.09, *SE*=0.17, *z*=-0.54, *p*=0.589); adults showed no significant difference (low variability, *M*=88%, *SD*=4%; high variability, *M*=89%, *SD*=5%; β=0.16, *SE*=0.26, *z*=0.61, *p*=0.542).

Follow-up Bayes Factor Analyses

The statistics reported provide evidence that 7-year-olds are more affected by variability than adults, and that only 7-year-olds show evidence of a low variability benefit during training. However, these frequentist statistics cannot inform us as to the likelihood of the null hypothesis, and we therefore conducted follow-up *BF* analyses to see whether there is evidence that (i) adults and (ii) 10-year-olds do *not* show a benefit for low variability items equivalent to that seen in 7-year-olds.

Our data summaries were taken from the coefficients reported above (10-year-olds: mean-diff=0.09, *SE*=0.17; adults: mean-diff=-0.16, *SE*=0.26). We base our estimate of the predicted mean difference on that of 7-year-olds, and thus set the *SD* of the half normal for H1 to 0.38. For 10-year-olds, $BF_{0,\,0.38}$=0.64, and for adults, $BF_{0,\,0.38}$= 0.39. We do not have substantial evidence for the null in either case.

Summary of Training Data

All age-groups improved throughout the training task and mean accuracy across the task

was high (above 80%) for each age-group. As predicted, adults outperformed both age-groups of children, and 10-year-olds outperformed 7-year-olds. Our key prediction (following Giannakopoulou et al., 2017) was that words heard from a single speaker would be identified more accurately than words heard from multiple speakers. In fact, the effect of variability interacted with age: there was a significant interaction with the contrast between adults and 7-year-olds; 10-year-olds did not significantly differ from either adults or younger children. Looking at each group separately, numerically, both groups of children showed performance in line with the prediction (better performance for low variability), whilst adults showed a reversed benefit for high variability. However, the difference between conditions was only significant for 7-year-olds. *BF* analyses indicated that there was not substantial evidence for the null for either 10-year-olds or adults.

Production Test

Production data were binary coded as correct/incorrect by comparing the transcriptions of participant productions with the correct responses. This differs from the approach taken by Barcroft and Sommers (2005), who coded partial errors using 0.5 and conducted ANOVA over averages. We use a binary coding scheme so as to be able to use logistic regression and avoid non-normality[4]. Responses were coded as correct if they contained an approximation of each of the target phonemes in the correct order. Cases where participants substituted the closest English

[4] Coding percent phoneme correct as a continuous measure leads to data which is highly skewed and does not meet the assumptions of normality required of linear mixed effect models.

equivalent for Lithuanian phonemes (e.g., used a voiced alveolar approximant for the trill /r/ or got the vowel length wrong but the quality correct) were counted as correct. Each word was tested twice giving participants two opportunities to show their knowledge of a word and potentially reducing floor effects in children. There were therefore two data points per participant and the maximum possible score for Production test was 24.The proportion of correct productions split by *variability-condition* for each *age-group* is shown in Figure 4.

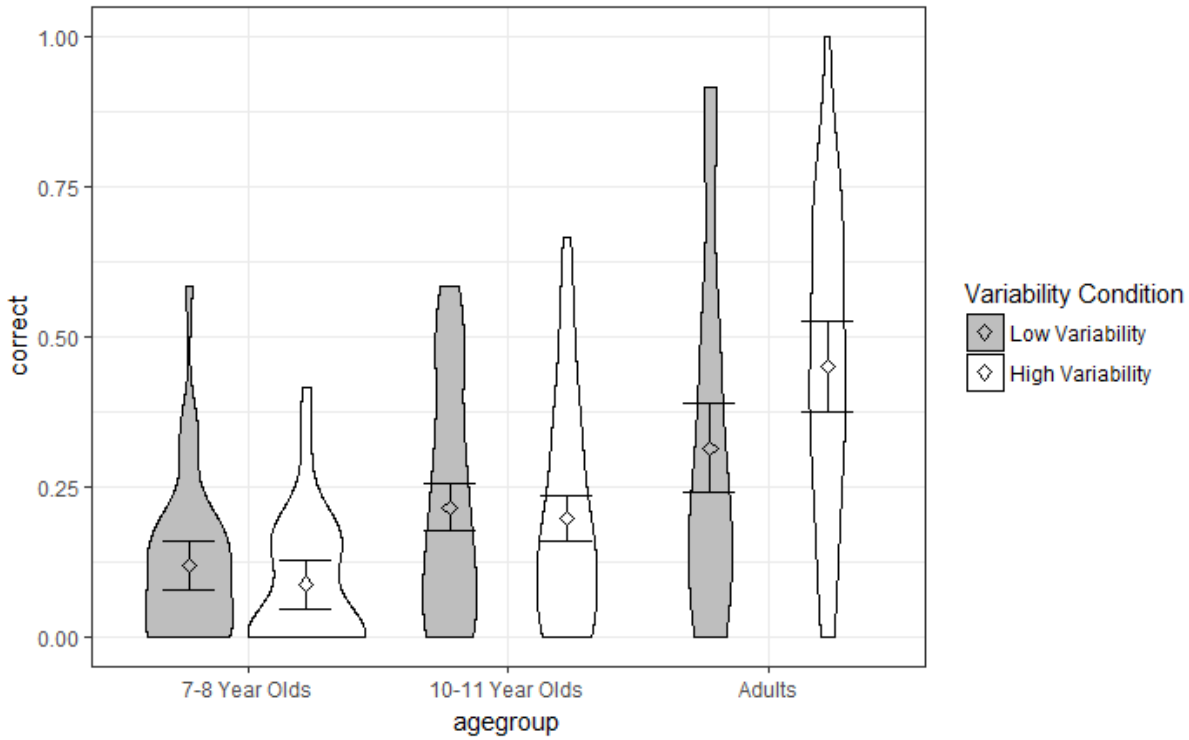


Figure 4: *Violin plot showing the proportion of correct response in the Production test for high variability and low variability items. Shape shows the kernel probability density of participants' mean scores. Mean values are shown; error bars indicate 95% confidence intervals around the*

mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Morey, 2008).

Structure of the Glmer Model

Predicted Outcome: Production accuracy (1/0).

Fixed Factors: *Variability-condition* and the contrasts for *age-group,* and their interaction, coded as in the model of training data.

Random Effects: Random intercepts for *participant* and *word* were included. As with the training data, we automatically included full random slopes for participant (i.e., a slope for *variability-condition* and the correlations between the intercept and the slope) and worked backwards to find the maximum slope structure supported by the model for *word.* The final model included by-word slopes for *variability-condition* and *age-group,* but not for the interaction between them and with no correlations between slopes.

Results in Final Model

Although performance in this task was lower than in training and comprehension, as predicted, adults outperformed both 10-year-olds and 7-year-olds, and 10-year-olds outperformed 7-year-olds (7-year-olds, M=10%, SD=9%; 10-year-olds, M=21%, SD=17%; adults, M=38%, SD=23%); 10-year-olds vs. adults: β=1.4, SE=0.43, z=3.28, p=.001; 7-year-olds vs. adults: β=2.53, SE=0.5, z=5.07, p<.001; 7-year-olds vs. 10-year-olds: β=1.14, SE=0.54, z=2.1, p=.036). Overall, there was no main effect of variability (β=0.13, SE=0.28, z=0.47, p=.64) however there was a significant interaction between variability and both the contrast between 10-

year-olds and adults (β=1.08, *SE*=0.42, *z*=2.59, *p*=.01) and the contrast between 7-year-olds and adults (β=1.53, *SE*=0.47, *z*=3.28, *p*=.001), though there was no interaction between variability and the contrast between 7-year-olds and 10-year-olds (β=0.45, *SE*=0.46, *z*=0.97, *p*=.331). Separate slopes for variability for each age-group indicated that adults performed significantly better with high (*M*=45%, *SD*=25%) compared to low (*M*=32%, *SD*=29%) variability items (β=1, *SE*=0.34, *z*=2.91, *p*=.004) but that there was no significant difference between the variability conditions for either 7-year-olds (high variability, *M*=9%, *SD*=11%; low variability, *M*=12%, *SD*=13%; β=-0.53, *SE*=0.42, *z*=-1.26, *p*=.207) or 10-year-olds (high variability, *M*=20%, *SD*=19%; low variability, *M*=22%, *SD*=19%; β=-0.08, *SE*=0.37, *z*=-0.22, *p*=.825).

Follow-up Bayes Factor Analyses

BF analyses further investigated whether there is evidence that (i) 7-year-olds and (ii) 10-year-olds do *not* show a benefit for high variability items equivalent to that seen in adults. Our data summary for each age-group is again taken from the coefficients reported above (i.e., 7-year-olds: mean-diff=0.60, *SE*=0.43; 10-year-olds: mean-diff=-0.13, *SE*=0.38); the *SD* of the half normal for H1 was set to 0.90 (corresponding to an odds ratio of 2.26). For 10-year-olds, the $BF_{0,\,0.90}$=0.29 and for 7-year-olds, the $BF_{0,\,0.90}$=0.18. Following the conventions in Dienes (2008), this suggests substantial evidence for the null in both age-groups.[5]

[5] Dienes (pc) also recommends reporting a robustness region – that is, the range of values of H1 (used to set *SD* of the theory) where there is substantial evidence for the null. For 7-year-olds, the null holds for values above 0.52 (odds ratio of 1.68); for 10-year-olds, the null holds for values above 0.88 (odds ratio of 2.25).

Summary of Production Test Data

Although performance was low, all age-groups showed some degree of word learning. As expected, adults outperformed children, and 10-year-olds outperformed 7-year-olds. Our critical prediction was higher performance with high variability items. In fact, the effect of variability interacted with age: specifically, there was a significant interaction with both the contrast between 7-year-olds and adults, and the contrast between 10-year-olds and adults. Looking at each group separately, adults showed the predicted significantly higher performance in the high variability condition; neither 7-year-olds nor 10-year-olds showed any significant difference between variability conditions. *BF* analyses found substantial evidence for the null for both child age-groups.

Comprehension Test

Responses in the comprehension test were coded as correct/incorrect. The proportion of correct responses, split by *variability-condition* for each *age-group* is shown in Figure 5.

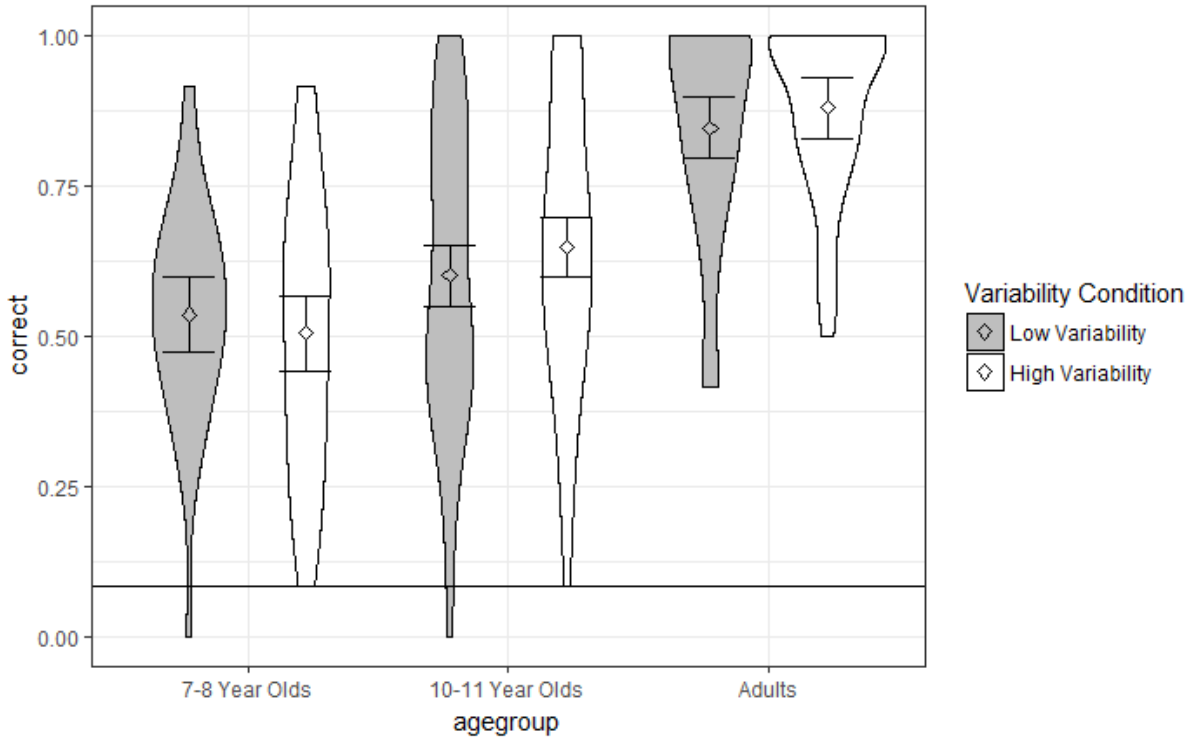


Figure 5: *Violin plot showing the proportion of correct response in the Comprehension test for high and low variability items. Shape shows the kernel probability density of participants' mean scores. Mean values are shown; error bars indicate 95% confidence intervals around the mean after between-subject variability has been removed, which is appropriate for repeated-measures comparisons (Morey, 2008). Horizontal line indicates chance (i.e., 1/12).*

Structure of the Glmer Model[6]

Predicted outcome: Correct picture selection (1/0).

Fixed factors: *Variability-condition* and the contrasts for *age-group,* and their interaction, coded as for the training and production data. We also tested whether adding *speaker* as a fixed effect improved the model fit – it did not (χ^2=0.078, p=.78) and so was not included (in this task speaker should not be included as a random effect as there are only two speakers; we do not automatically include it as a fixed effect as it was not an experimental factor).

Random effects: Random intercepts for *participant* and *word* were included. Full random slopes were included for participant (i.e., a slope for *variability-condition* and the correlations between the intercept and the slope) and worked backwards to find the maximum slope structure supported by the model for *word.* The final model included by-word slopes for *variability-condition,* and *age-group* but not the interaction between them, and no correlation between slopes.

Results in Final Model

Although participant groups were well above chance in this task, as predicted, adults outperformed both 10-year-olds and 7-year-olds, and 10-year-olds outperformed 7-year-olds (7-year-olds, M=53%, SD=17%; 10-year-olds, M=63%, SD=23%; adults, M=86%, SD=13%; 10-year-olds vs. adults: β=1.94, SE=0.41, z=4.72, p<.001; 7-year-olds vs. adults: β=2.61, SE=0.39,

[6] Due to a computer error, one participant received 15 additional trials. These were removed from their data set before analyses were conducted.

z=6.6, p<.001; 7-year-olds vs. 10-year-olds: β=0.67, SE=0.33, z=2.03, p=.042). Overall, there was no main effect of *variability-condition* (β=0.18, SE=0.18, z=0.98, p=.326). There was no interaction between *variability-condition* and the contrasts between *age-groups* (10-year-olds vs. adults: β=0.09, SE=0.57, z=0.15, p=.879; 7-year-olds vs. adults β=0.55, SE=0.48, z=1.14, p=.254; 7-year-olds vs. 10-year-olds: β=0.47, SE=0.43, z=1.07, p=.283).

Follow-up Bayes Factor Analyses

Again we performed BF analyses to investigate whether we have evidence for the null hypothesis (i.e., no high variability benefit) for each age-group. We used the same method as previously; however, since none of the age-groups showed a significant variability benefit, it is less clear how to choose a suitable value to inform H1. Since we have no relevant value for this particular test, we continue to use the estimate from the adult production data, i.e., 0.9 (odds ratio 2.46), on the assumption that a roughly similar sized difference would be predicted across production and comprehension. Our data from each group is the mean-diff and *SE* from a version of the model with a separate slope fitted for variability for each age-group (i.e., 7-year-olds: β=-0.16, SE=0.28, 10-year-olds: β=0.31, SE=0.34, adults: β=0.39, SE=0.39). Bayes factors were: 7-year-olds, $BF_{0,\,0.90}$=0.18; 10-year-olds, $BF_{0,\,0.90}$=0.74; adults, $BF_{0,\,0.90}$=0.92. Following the conventions in Dienes (2008), we have substantial evidence for the null only for the 7-year-olds.[7]

[7] As before, we computed the robustness region, that is, range of values of H1 (used to set *SD* of the theory) where there is substantial evidence for the null. For 7-year-olds, the null holds for values above 0.51 (i.e., an odds ratio of ~1.67). Note that an alternative more conservative method to inform H1 would be to base it on mean difference from

Summary of Comprehension Test Data

Performance was generally high in all groups, and as predicted, adults outperformed both groups of children, and 10-year-olds outperformed 7-year-olds. Contrary to our key prediction there was no evidence of higher accuracy with high variability items for any of the age-groups, and there were no interactions between *variability-condition* and *age-group*. *BF* analyses found substantial evidence for the null over the theory for 7-year-olds but not for either adults or 10-year-olds, where there was no substantial evidence either for the theory or the null.

General Discussion

The current study investigated speaker variability effects in L2 word learning in 7-8 year-old children, 10-11 year-old children and adults. English speaking participants learned L2 Lithuanian words using a 2AFC picture selection task in which half of the words were spoken by a single speaker (low variability) and the other half were spoken by eight speakers (high variability). Training performance was recorded as well as performance on follow up production (picture-to-word recall) and comprehension (word-to-picture recall) tests. In all tasks, adults outperformed children, and 10-year-olds olds outperformed 7-year-olds, in line with a benefit of age seen in previous studies of vocabulary learning (e.g., Henderson et al., 2013). However, our

chance for this age group (in log odds space). This would give a value of 2.52, that is, within the range of substantial values.

key predictions concerned the effects of speaker variability. First, following Giannakopoulou et al. (2017), we expected to see higher performance on the one-speaker items (a low variability benefit) during the training task. In fact, there was an interaction with age-group, with substantial evidence for this difference only in the 7-year-olds. Second, following Barcroft and Sommers (2005, 2014; Sommers & Barcroft 2007, 2011 – adult L2 learning) and Richstmeier et al. (2009 – child L1 learning), in both the production and comprehension tests, we predicted higher accuracy for the words learnt from multiple speakers (a high-variability benefit). For adults, this hypothesis was confirmed, but only in the production test; in the comprehension test, the variability effect was inconclusive (i.e., BF>.3 and <3), and results were close to ceiling. In contrast, 7-year-olds showed substantial evidence for the null (no variability benefit) in both production and comprehension; 10-year-olds showed substantial evidence for the null in production, with ambiguous evidence in comprehension.

How do these findings sit with the previous literature showing evidence for a benefit of multiple-speaker input in vocabulary learning? A clear point of convergence is the adult production data, where we saw clear evidence for a benefit of speaker variability, such that adults accurately recalled more of the words which they had heard exemplified by multiple speakers. This is in line with the findings of Barcroft and Sommers (2005) and extends their result to a different training paradigm (2AFC picture identification), as well as to a new language. Taken together, these findings provide evidence that – at least in the initial stages of acquiring novel word forms – indexical information present in the input affects the nature of the lexical representations which are formed. This is consistent with a considerable body of research

which speaks against any account in which speaker variation is normalized independently of the formation of lexical representations (Newman 2008; Creel & Bregman 2011).

Given the clear result from our adult production test, it is at first glance surprising that we did *not* find a corresponding variability benefit in the adult comprehension test, since this has also been found in earlier research. It is important to recognize, however, that the *BF* (.9) indicates that we do *not* have clear evidence for the null here – rather the data are insensitive. It thus is possible that with increased power we might obtain evidence for a variability benefit. However – unlike in production – adults' overall performance on the comprehension test is very high, with many participants at ceiling on both item types. This contrasts with performance in the equivalent tests in Barcroft, Sommers and colleagues' studies, and is undoubtedly due to the changes we made to our paradigm in order to be able to use the same materials with children – that is, reducing the number of items to be learned, increasing the number of repetitions per item during training, and using a picture identification test rather than a translation test. This highlights the methodological difficulty of conducting "matched" experiments with different age-groups; future work using this paradigm could perhaps incorporate measures of reaction times at test (though these would most likely be less useful with child participants).

Turning now to children, we did *not* see a variability benefit in either Production or Comprehension, for either age-group, with *BF*s showing substantial evidence for the null in both tests for 7-year-olds, and for Production in 10-year-olds. This contrasts with the findings of Richtsmeier et al. (2009), who *did* see a benefit in their L1 vocabulary experiment with 4-year-olds, and Rost and McMurray (2009, 2010) who saw a similar benefit with 18-month-olds,

although it is in line with the findings of Giannakopolou et al. (2017) in their L2 training experiment with 7-year-olds. We consider potential explanations for the lack of variability benefit for each age-group in turn.

Beginning with 7-year-olds, we first note that a potential concern is that there are floor effects in Production (average accuracy 10%). This could mask potential differences between conditions. However, we note that performance is *not* at or near floor in Comprehension (53%, where chance = 8.3%), and we again see substantial evidence for the null in this test. We therefore require a further explanation of why we do *not* see a variability benefit at least in this test. We suggest that the data collected during the training portion of the experiment provides an insight as to why this age-group does not show a high variability benefit. Specifically, 7-year-olds had greater difficulty identifying the correct referents for the target words which they encountered from multiple speakers than they did for the words produced by a single speaker. This is likely due to the fact that when there is one speaker exemplifying the target words, participants are required to attune to only one set of idiosyncratic speaker features for those items, leading to these items being identified more quickly over repeated trials. In previous work, greater ease of adapting to a single talker has also been seen in adults, in both L2 learning and L1 speech processing (Clopper & Pisoni, 2004; Giannakopoulou et al., 2017; Martin et al., 1989; Mullenix et al., 1989; Nusbaum & Morin, 1992). In the current paradigm, the data from adults during training is ambiguous (we do not have substantial evidence for either a low variability benefit, or for the null) but a significant interaction indicates that any low variability benefit is at least larger in 7-year-olds. This is consistent with the results of a study comparing the effects of

talker variability in L1 word recognition in younger children (aged 3, 4 and 5) which showed benefits of a single-talker, but that processing of multiple talkers becoming easier with age (Ryalls & Pisoni, 1997). One possibility is that age difference in processing multi-talker input are due to differences in working memory capacity. Consistent with this, Nusbaum and Morin (1992) found that – in adults – high variability stimuli place a particular burden on a learner's working memory. Although we do not have direct measures of verbal working memory in the current paper, it is well established that this increases through childhood (Case, Kurland & Goldberg, 1982; Alloway, Gathercole & Pickering, 2006). Our 7-year-olds may thus particularly struggle with the high variability stimuli during training.

In light of these findings, we suggest that the greater difficulty which 7-year-olds experience when processing the multiple speaker words during training makes it more difficult for them to make use of the multiple indexical features which – in adults – lead to the formation of more robust representations. That is, the potential benefits associated with encoding speaker information for subsequent retrieval are likely to be outweighed by the processing cost in this age group. This leads to no benefit for those words at test. At first glance, an explanation in terms of age-related processing difficulties might seem at odds with the findings of Richstmeier et al. (2009) and Rost and McMurray (2009, 2010), given that they found a benefit of high speaker variability in much younger children than those in the current study. However, recall that, in contrast to the current paradigm, the stimuli in both of those studies were native language non-words. One possibility is that it is the greater difficulty of dealing with multiple talker cues in the context of unfamiliar phonology which particularly impacts on children's ability to benefit

from these cues. A follow-up experiment using the same approach but using stimuli produced by native English speakers who pronounced the target Lithuanian words with English phonology could potentially shed light on the extent to which the greater difficulty in children L2 word learning arises from the unfamiliar non-native phonology. Other aspects of the experimental paradigm may also interact with the task complexity. For example, although the children were much younger, the task in Rost and McMurray (2009, 2010) was considerably easier (i.e., teaching children just two words and testing using looking time measures). Further evidence that task difficulty interacts with variability comes from a study by Goldinger, Pisoni & Logan (1991), which found that presentation rates affected the recall of single versus multi-talker word lists (in L1): single-talker lists produced better word recall than multiple-talker lists at short inter-word intervals (less than 2000 ms) whereas this effect was reversed for longer inter-word intervals. The authors interpret this in terms of the time needed to encode indexical properties. Importantly, it shows that, even for adults remembering L1 words, the nature of the task may place boundaries on the benefits of speaker variability.

Turning to our 10-11 year-old learners, in general this group showed a middle performance between adults and 7-year-olds, both in terms of overall performance and in terms of the patterns seen with respect to variability. As for adults, results from the comprehension test were ambiguous (no substantial evidence of a high variability benefit, but also no evidence for the null). However, unlike with adults, this does not appear to be due to ceiling effects, suggesting that a more conclusive result could be obtained with a larger sample. (We note that the adult Barcroft and Sommers study had a larger sample (N=60), compared to the current study

(N=32 per age-group), although the Richstmeier et al. (2009) study with child learners used similar sample sizes to the current study (N=43).) However, results were clear in production; just as was found for 7-year-olds, there was substantial evidence for the null – that is, no high variability benefit. We tentatively suggest that the absence of a variability benefit is again due to the greater difficulty that 10-year-olds have in processing the multiple speaker items during training compared with adults, although we acknowledge that we do not have direct statistical evidence for this in the data (since evidence from the training task was again ambiguous for this age-group).

Taken together, our results support the claim that "irrelevant" speaker cues *can* aid the early stages of lexical learning, but suggest that this may be constrained by task difficulty and learner capacity. In terms of the theories proposed in the literature, our results speak directly against an account in which the benefit of variability is itself a result of the difficulty of encoding these items, with greater effort itself leading to better later retrieval (see Barcroft, 2001, for discussion of this "cognitive effort" hypothesis). That hypothesis would predict that in the current study younger learners would show a *greater* variability benefit than adults, which is clearly not the case. This concurs with the finding that increasing encoding difficulty in adults (by decreasing the signal to noise ratio) does *not* lead to benefit like that seen for high variability items (Barcroft, 2001). Instead, the results are in line with an account in which indexical cues are somehow incorporated into lexical representations, as suggested by the *elaborative processing* hypothesis (Barcroft, 2001). They are also consistent with the discriminative learning account discussed in the introduction in which the presence of uninformative speaker-specific features

assists learning by decontextualizing lexical codes (Ramscar & Baayen, 2013; Ramscar et al., 2010).

Our results suggest there may be boundaries on the benefit of speaker variability (cf. Goldinger et al., 1991). More generally, a key "take home" of this work is that manipulations which increase the complexity of the input may have different effects on learners at different ages. It therefore cannot be assumed that materials which are more effective for adult learners will necessarily be so for children. Looking outside of the literature on lexical learning, our results are consistent with a broader literature on how a learner's cognitive abilities might constrain their ability to benefit from input variability. For example, for both syntactic and morphological learning, there is both computational and empirical evidence that encountering grammatical morphemes across a broader range of vocabulary – that is, *lexical* variability – promotes generalisation (Bybee, 1995; Gomez, 2002; Plunkett & Marchman, 1991; Wonnacott, Boyd, Thomson & Goldberg, 2012). Brooks, Kempe and Sionov (2006) explored this effect in English speaking adult learners exposed to an unfamiliar L2 (Russian). They found evidence that greater variability promoted generalisation *only* in learners with above median scores on an IQ test, which they interpret as showing a mediating role of executive function and attention resources. Our own ongoing work explores the extent to which such constraints are relevant in learners of different ages, and at different stages of learning, in a range of linguistic contexts.

Given that Richstmeier et al. (2009) found that even younger children than those tested here *did* show a benefit of speaker variability for native input, an important direction for future work will be to discover whether benefits of speaker variability might be present at a later stage

of child L2 learning, once the phonology becomes more familiar.

Our results have implications for the development of foreign language instruction materials at different ages. A potential limitation is that – as in the other vocabulary training studies reviewed in this paper – we use a somewhat artificial training paradigm, in which vocabulary is trained in isolation rather than encountered in context. We note that this is not necessarily problematic given that foreign language instruction does make use of such direct vocabulary instruction. Nevertheless, it is an open question whether the variability effect seen here (and elsewhere) in adult learners would apply in more naturalistic contexts, and whether the same types of constraints on children's learning would apply.

In conclusion, while the data from our adult learners (at least from production) supports a theory in which the presence of "irrelevant" speaker identity cues aid lexical learning, our child data suggests that this benefit is constrained by the relative difficulty of the learning task for the learner in question. Our use of Bayes factor analyses allowed us to demonstrate substantial evidence for the null hypothesis that there is no variability benefit for 10-year-olds in production and for 7-year-olds in both production and comprehension. We attribute these null results to children's greater difficulty in processing items produced by multiple speakers during training – for which we have direct evidence at least for the 7-year-olds. This greater processing cost appears to prevent idiosyncratic speaker cues from playing a beneficial role in lexical learning. This highlights that manipulations which benefit adult L2 learners cannot be assumed to apply equally to younger learners.

Acknowledgements

This research was supported by the Economic and Social Research Council (grant number: ES/K013637/1, awarded to EW and HB). We would like to thank the Lithuanian speakers who agreed to lend their voices for the stimuli preparation as well as Agne Sinkeviciute for pilot testing.

References

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

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

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

Barcroft, J. (2001). Acoustic variation and lexical acquisition. *Language Learning, 51,* 563-590. https://doi.org/10.1111/0023-8333.00168

Barcroft, J., & Sommers, M.S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27*, 387 – 414. https://doi.org/10.1017/s0272263105050175

Barcroft, J., & Sommers, M.S. (2014). Effects of variability in fundamental frequency on L2 vocabulary learning. *Studies in Second Language Acquisition, 36,* 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,*

44, 255-278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., & Bolker, B. (2013). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-0.

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.0.36, retrieved 11 November 2017 from http://www.praat.org/

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*, 977 – 985. https://doi.org/10.3758/bf03206911

Brooks, P.J., Kempe, V. & Sionov, A. (2006). The role of learner and input variables in learning inflectional morphology. *Applied Psycholinguistics, 27,* 185– 209. https://doi.org/10.1017/S0142716406060243

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes, 10*, 425-455. https://doi.org/10.1080/01690969508407111

Case, R., Kurland, D.M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. Journal of Experimental Child Psychology, 33, 386–404. doi:10.1016/0022-0965(82)90054-6

Children's Printed Word Database [online]. Available from: <http://www.essex.ac.uk/psychology/cpwd/> [Accessed October 8 2017].

Clopper, C.G., & Pisoni, D.B. (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech, 47*, 207-239.

https://doi.org/10.1177/00238309040470030101

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

DeKeyser, R.M. (2012). Age effects in second language learning. In *S. Gass & A. Mackey (Eds.), Handbook of Second Language Acquisition* (pp. 442-460). London: Routledge. https://doi.org/10.4324/9780203808184.ch27

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

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

Dienes, Z. (2015). How Bayesian statistics are needed to determine whether mental states are unconscious. In *M. Overgaard (Ed.), Behavioural Methods in Consciousness Research.* Oxford: Oxford University Press, pp 199-220. https://doi.org/10.1093/acprof:oso/9780199688890.003.0012

Galle, M., Apfelbaum, K., & McMurray, B. (2015). The role of single talker acoustic variation in early word learning. *Language Learning and Development, 11,* 66–79. https://doi.org/10.1080/15475441.2014.895249

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

Goldinger, S.D., Pisoni, D.B., & Logan, J.S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory and Cognition, 17(1),* 152-162. doi.org/10.1037//0278-7393.17.1.152

Gomez, R.L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436. https://doi.org/10.1111/1467-9280.00476

Henderson, L.M., Weighall, A., Brown, H., & Gaskell, M.G. (2013). On-line lexical competition during spoken word recognition and word learning in children and adults. *Child Development, 84,* 1668-1685. https://doi.org/10.1111/cdev.12067

Huang, B.H. (2016). A synthesis of empirical research on the linguistic outcomes of early foreign language instruction. *International Journal of Multilingualism, 13,* 257-273. https://doi.org/10.1080/14790718.2015.1066792

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*, 145-160. https://doi.org/10.1017/S0142716411000300

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

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,* 66-99. http://dx.doi.org/10.1016/0010-0285(89)90003-0

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for

30,000 English words. *Behavior Research Methods, 44,* 978–990. https://doi.org/10.3758/s13428-012-0210-4

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. *Journal of the Acoustical Society of America, 94,* 1242-1255. https://doi.org/10.1121/1.408177

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

Martin, C.S., Mullenix, 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,* 676-684. https://doi.org/10.1037//0278-7393.15.4.676

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, 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

Morey, R.D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology, 4,* 61-64. https://doi.org/10.20982/tqmp.04.2.p061

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

Newman, R. S. (2008). The level of detail in infants' word learning. *Current Directions in Psychological Science, 17*(3), 229-232. https://doi.org/10.1111%2Fj.1467-8721.2008.00580.x

Nusbaum, H.C., & Morin, T.M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech Perception, Speech Production, and Linguistic Structure*, 113-134. Tokyo: OHM.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition, 38,* 43-102. https://doi.org/10.1016/0010-0277(91)90022-V

Pruitt, J.S. (1993). Comments on 'Training Japanese listeners to identify /r/ and /l/: A first report' [J.S. Logan, S.E. Lively, and D B. Pisoni, *Journal of the Acoustical Society of America, 89,* 874-886 (1991)], *Journal of the Acoustical Society of America, 94,* 1146-1147. https://doi.org/10.1121/1.406962

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

R Core Team. (2016). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Available at https://www.R-project.org/.

Ramscar, M. and Baayen, R.H. (2013). Production, comprehension, and synthesis: A communicative perspective on language. Frontiers in Language Sciences. https://doi.org/10.3389/fpsyg.2013.00233

Ramscar, M., Yarlett, D., Dye, M., Denny, K., and Thorpe, K. (2010). The effects of feature label-order and their implications for symbolic learning. *Cognitive Science, 34,* 909–957. https://doi.org/10.1111/j.1551-6709.2009.01092.x

Richtsmeier, P.T., Gerken, L., Goffman, L., & Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition, 111,* 372-377. https://doi.org/10.1016/j.cognition.2009.02.009

Rost, G.C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*, 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,* 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Ryalls, B.O., & Pisoni, D.B. (1997). The effect of talker variability on word recognition in preschool children. Dev Psychol, 33(3), 441–452. doi.org/10.1037//0012-1649.33.3.441

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

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

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,* 231 – 249. https://doi.org/10.1017/s0142716407070129

Sommers, M.S., & Barcroft, J. (2011). Indexical information, encoding difficulty, and second language vocabulary learning. Evidence from amplitude, fundamental frequency, and speaking rate variability. *Applied Psycholinguistics, 32,* 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,* 381–382. https://doi.org/10.1038/41102

Statistical Office of the European Communities. (2015). EUROSTAT: Regional statistics: Foreign Language Learning. Luxembourg: Eurostat.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics, 36,* 131 – 145. https://doi.org/10.3758/bf03202673

Wonnacott, E., Boyd, J.K., Thomson, J., & Goldberg, A.E. (2012). Input effects on the acquisition of a novel phrasal construction in 5 year olds. *Journal of Memory and Language, 66,* 458-478. https://doi.org/10.1016/j.jml.2011.11.004

Wells, J.C., (1997). 'SAMPA computer readable phonetic alphabet'. In Gibbon, D., Moore, R. and Winski, R. (Eds.), 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin and New York: Mouton de Gruyter. Part IV, section B.

Werker J.F., & Curtin S. (2005). PRIMIR: a developmental framework of infant speech

processing. *Language Learning and Development*, *1*, 197–234.

https://doi.org/10.1080/15475441.2005.9684216

Appendix A

Table 2: *Native languages spoken by bilingual participants*

7-8 year-olds		10-11 year-olds		Adults	
Language	**Number of participants**	**Language**	**Number of participants**	**Language**	**Number of participants**
Bengali	1	Albanian	1	Bengali	1
French	1	Arabic	1	German	1
German	1	Filipino	1	Jamaican Patois	1
Hungarian	1	Somali	1	Polish	1
Jamaican Patois	1	Bengali	2	Punjabi	1
Malayalam	1	German	2	Thai	1
Portuguese	1	Pashto	2	Urdu	1
Albanian	2			Greek	2
Spanish	2			Cantonese	3
Polish	3			Hindi	3
Twi	3			Mandarin	3