Applied Psycholinguistics (2024), 1–24
doi:10.1017/S0142716424000109

ORIGINAL ARTICLE

Effects of speech production training on memory across short and long delays in 5- and 6-year-olds: A pre-registered study

Belén López Assef[1], Margarethe McDonald[1,2], Amélie Bernard[1,2] and Tania S. Zamuner[1]

[1]Department of Linguistics, University of Ottawa, Ottawa, ON, Canada and [2]School of Psychology, University of Ottawa, Ottawa, ON, Canada
Corresponding author: Belén López Assef; Email: mlope075@uottawa.ca

(Received 1 May 2023; revised 28 January 2024; accepted 23 February 2024)

Abstract

Studies on the role of speech production on learning have found a memory benefit from production labeled the "Production Effect." While research with adults has generally shown a robust memory advantage for produced words, children show more mixed results, and the advantage is affected by age, cognitive, and linguistic factors. With adults, the Production Effect is not restricted to the immediate context but is also found after a delay. So far, no studies have investigated the effect of delayed recall on the Production Effect with children. Children aged 5 and 6 years old ($n = 60$) participated in two sessions. Children were trained on familiar words and images, which were heard (*Listen*) or produced aloud (*Say*). Children then performed a free recall task. One week later, children repeated the recall task and an additional recognition task. At immediate testing, there was a recency effect on words recalled from the different training conditions and a recall advantage for words produced over words heard; however, this no longer held after a 1-week delay in either the recall or recognition task. Exploratory analysis showed that vocabulary did not predict the Production Effect. Findings indicate that unlike adults, the Production Effect is not as robust in children after a delay.

Keywords: delayed testing; long-term memory; production effect; recall; speech production

For years, researchers from many areas have been interested in how different factors could potentially help or improve memorization and learning. Actions have been found to be helpful during learning, for example, pairing gestures with new words that were heard during learning improves later recognition (Mayer et al., 2015). Another action that has been identified as potentially beneficial is speech production: saying words aloud. The intuition that saying words aloud will somehow help memorization and learning has been confirmed by a large body of research dedicated to what is labeled as "The Production Effect" (term coined by

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

MacLeod et al., 2010). The Production Effect refers to a memory advantage for produced items compared to items that have been only heard or only seen. Previous studies have consistently found that the Production Effect generalizes across different conditions or populations, such as different age groups (children, Icht & Mama, 2015; adults, MacLeod et al., 2010), language familiarity (L1, Bodner & Taikh, 2012; L2, Icht & Mama, 2019), testing methodologies (free recall task, Cho & Feldman, 2016; recognition task, Zamuner et al., 2016; translation task, Kaushanskaya & Yoo, 2011), and stimuli characteristics (images, Icht et al., 2020; auditory stimuli, Forrin & MacLeod, 2016).

Since the Production Effect has been presented as a memory tool for children (Icht & Mama, 2015), a key question to ask is whether the Production Effect is restricted to the immediate experimental setting or whether this memory advantage persists over time. While studies with adults have found that the Production Effect persists after a delay in adults (1 and 2 weeks) (Icht & Mama, 2019; Kaushanskaya & Yoo, 2011; Ozubko et al., 2012), to our knowledge, no studies to date have investigated this with children. Delay effects in children, as opposed to adults, are of special interest as the Production Effect in children has been argued to vary depending on many factors, including age and task difficulty (Zamuner et al., 2017). Investigating the effect of a delay on the Production Effect in children not only informs our theories on the Production Effect and the multiple factors that possibly affect it but also informs theories about long-term memory and encoding in children, along with potentially informing psycholinguistic theories on the relationship between perception and production. To further our understanding of the Production Effect in long-term memory, the current study extends previous research with children by looking at whether the production advantage is maintained after a 1-week delay. The study's methodology, hypotheses, and analysis plan were pre-registered, see https://osf.io/3guay/?view_only=eb042fdb21874e35913f37465acbbcd4. The study was divided into 2 sessions: Immediate testing and Delayed testing (1-week later). Children aged 5 and 6 years were trained on images of familiar objects in two conditions: *Listen* (participants see the image and then hear a recording of the corresponding word) and *Say* (participants see the image and say the corresponding word aloud). After training, children were tested on a free recall task. One week later, participants were tested again on the free recall task, and they also completed an additional Old/New recognition test. We predicted that if the Production Effect creates distinct encodings for items that are being memorized, the memory trace for *Say* items would be stronger than for *Listen* items. Moreover, we predicted that if this distinctiveness is not susceptible to children's less-developed cognitive skills, this would lead to better recall for *Say* over *Listen* words at both immediate and delayed testing, similar to the effect seen with adults.

The production effect: adults

The Production Effect was first proposed by MacLeod et al. (2010). They found a memory advantage for items that were read aloud over items that were read silently in a series of experiments. The foundational work by MacLeod and colleagues built upon previous research that had found an advantage for production when

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

compared to silent reading and mouthing (Conway & Gathercole, 1987; Gathercole & Conway, 1988; Hopkins & Edwards, 1972). MacLeod and colleagues proposed distinctiveness as the reason behind the advantage: produced items are more distinctive than read-only items and this distinctive feature is encoded in memory and used when performing a recall or recognition task.

While many studies have shown a robust memory advantage for produced items (e.g., Forrin et al., 2012; Hopkins & Edwards, 1972; Icht et al., 2020; MacLeod et al., 2010; Ozubko & MacLeod, 2010), some studies have found the opposite: an attenuation of the effect or a disadvantage for produced items. This change in direction of the effect for produced items arises when various factors are manipulated, such as the level of familiarity with the stimuli being used in the study (Baese-Berk, 2019; Baese-Berk & Samuel, 2016; Cho & Feldman, 2016; Kaushanskaya & Yoo, 2011, López Assef et al. 2023) and design of the study (list discrimination task, Bodner & Taikh, 2012; between- vs. within-subjects, Jones & Pyc, 2014; mixed vs. pure lists, Ozbuko & MacLeod, 2010). Another factor that has been found to interact with the Production Effect is the order in which the stimuli are presented (Cyr et al., 2021; Gionet et al., 2022; Saint-Aubin et al., 2021). For example, Saint-Aubin et al. (2021) found that items read silently showed better recall when they appeared earlier in the experiment, compared to produced items, which showed better recall when they appeared towards the end.

Furthermore, the Production Effect is also impacted by the amount of training: while participants might benefit from production initially, increasing the number of production training trials (and thus, increasing the number of repetitions per item) can reverse the positive effect from speech production when learning novel words (Kapnoula & Samuel, 2022). Timing of production seems to also affect the Production Effect. Studies have found learning disruptions when producing non-native sound contrasts (Baese-Berk & Samuel, 2022) and new words (Kapnoula & Samuel, 2023) immediately after exposure, but this disruption can be reduced by including a 4s delay between perception and production during the training trials (Baese-Berk & Samuel, 2022; Kapnoula & Samuel, 2023).

Delayed testing: adults

Previous research has investigated the effect of time on the Production Effect with adults (Grohe & Weber, 2018; Icht & Mama, 2019; Kaushanskaya & Yoo, 2011; Ozubko et al., 2012). Overall, the Production Effect has been found to be present in both immediate and delayed testing. There are lasting effects for the Production Effect in adults for different types of stimuli and different lengths of delays: the Production Effect is found for known words at immediate testing, after a 1-day delay and 1-week delay (Ozubko et al., 2012); for words in familiar and unfamiliar accented speech at both immediate testing and a 1-week delay (Grohe & Weber, 2018); in a second language vocabulary learning task at immediate testing, after 1 week and 2 weeks (Icht & Mama, 2019); and for novel words that were phonologically familiar at immediate testing and after 1 week (Kaushanskaya & Yoo, 2011).

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

The production effect: children

Currently, only four studies have investigated the Production Effect with children: Icht and Mama (2015), who tested 5-year-old children on familiar and unfamiliar objects (using real words but less frequent and not familiar to children); Zamuner et al. (2018), who tested 4.5- to 6-year-old children using eye-tracking and novel word learning; Pritchard et al. (2019), who tested 7- to 10-year-old children on familiar and novel words using a reading paradigm; and López Assef et al. (2021), who looked at the Production Effect with familiar objects across development in children 2 to 6 years old.

Some of these studies observed a Production Effect in children. Pritchard et al. (2019) found an advantage for produced words and novel words when comparing reading aloud to reading silently. Icht and Mama (2015) also found an advantage for produced items using familiar objects when compared to a "*Look*" condition (silently observing the picture) and "*Look and Listen*" condition (looking at the picture and hearing the experimenter say the corresponding word). Icht and Mama described their results in reference to the number of encoding processes for each item: the higher the number of encoding processes involved, the more likely it is to be remembered. More specifically, "*Look,*" silently observing the picture, involved only one encoding process: visual; "*Look and Listen,*" looking at the picture and hearing the experimenter say the corresponding word, involved in two encoding processes: visual and auditory; lastly "*Look and Say,*" looking at the picture and saying the corresponding word aloud, was defined as implicating three encoding processes (visual, auditory, and articulatory). Consistent with Icht and Mama's hypotheses, children's recall rates showed a gradient-like pattern, paralleling the number of encoding processes during training: recall was the lowest for "*Look,*" followed by "*Look and Listen,*" and "*Look and Say*" showed the highest recall. Thus, the memory benefit for produced words was attributed to increased distinctiveness, stemming from the higher number of encoding processes for produced items compared to "*Look*" or "*Look and Listen*" items. Icht and Mama argued that the greater number of encoding processes resulted in better memory for produced items because children had more information to use during activation and retrieval of the items from memory.

However, as with adults, other studies with children have also shown a reversal of the Production Effect. This was found when testing children on their recognition of novel words (Zamuner et al., 2018) and in younger children with familiar words (López Assef et al. 2021). López Assef and colleagues, using the same paradigm as Icht and Mama (2015) and the current paper, found a reversed Production Effect (advantage for heard items over produced items) in their younger children (2 to 3 years old) and a typical Production Effect for older children (5 to 6 years old). While production did not result in a memory advantage for younger children, as children grow older, their cognitive skills develop, thus allowing them to make use of the extra information provided by production.

Developmental differences in memory

In the discussion up until now, we have reviewed the findings where children were tested on recall immediately after training or exposure. In contrast with the adult literature, no studies to date have investigated the production effect and delayed

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

testing with children. Thus, we will inform our predictions for delayed testing and the production effect with children based on the existing literature looking at the effect of delayed testing on word learning with children.

Previous research has shown that children can remember information across different delays: 5-minutes (Sakhon et al., 2018), 1-hour (Wang et al., 2018), 1-week (Holland et al., 2015; Scarf et al., 2013), 1-month (Markson & Bloom, 1997), and multiple months (Kan, 2014; Wang et al., 2015), with memory usually declining as the delay increases (i.e., Lawson & London, 2015). While many studies suggest that children's memory peaks at immediate testing, and then slowly decreases after time, there are also instances where children show better memory at delayed testing. In a word-learning experiment, 3- to 5-year-olds showed a higher than chance of looking at a target image after a 1-week delay, but not at immediate testing or after a 5-minute delay (Sakhon et al., 2018). Similarly, when asking children about new information they learned, 5-year-olds showed better memory (higher accuracy) after a 2-to-3-day delay compared to immediate testing, whereas 4-year-olds showed similar performance at both immediate testing and after a 2-to-3-day delay (Bemis & Leichtman, 2019).

The likelihood of remembering something can be increased by engaging with the item either at the moment of encoding or afterward. For example, the inclusion of memory cues (like mnemonic devices) or support techniques (like multiple exposures to items) can increase 3-year-olds' retention of newly learnt words. Children who studied new words with support techniques showed better retention after initial exposure and were less likely to forget the words over time than children who learned without support techniques (Vlach & Sandhofer, 2012). Distinctive processing has been shown to play an important role from the very early stages of development, regardless of cognitive limitations, as it helps guide attention, perception, and encoding by making certain items stand out compared to others (Howe, 2006; Howe et al., 2000). Regarding the Production Effect, limitations in children's working memory and cognitive skills could affect the encoding process of the distinctive item, disturbing the Production Effect. For example, while Icht and Mama's study with 5-year-olds found that increased distinctiveness (more processing levels) helped memorization, it is possible that this memory benefit is disrupted in time, thus resulting in the Production Effect being present at immediate test, but not after a delay.

The current project extends investigation of the Production Effect in children by looking at whether the effect persists after a 1-week delay. That is, assuming children can create distinctive encodings for produced words, is this information still available after 1 week? Children aged 5 and 6 years old were trained on familiar words paired with pictures in two conditions: half of the words were said aloud by participants during training (*Say*) while for the other half, participants heard an audio recording of the word (*Listen*). Their memory of these words was tested using a free recall task immediately after training and after a 1-week delay. Testing after the delay also included a recognition task and a standardized vocabulary test.

Based on previous research with the same-aged children using familiar words (Icht & Mama, 2015; López Assef et al., 2021), we expected the Production Effect to be present at immediate testing. In other words, at immediate testing, there would be a memory advantage for items that were said aloud during training over items that were heard during training. We also had two possible hypotheses for how the

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

delay could affect the Production Effect. The first hypothesis was that the Production Effect creates distinct encodings for items that are being memorized, and these are not affected by children's less-developed cognitive skills. Thus, we expected that the memory trace for produced aloud items would be stronger than for heard-only items at both immediate and delayed testing. This follows from previous patterns observed in the literature from the Production Effect with delayed testing with adults (Grohe & Weber, 2018; Icht & Mama, 2019; Kaushanskaya & Yoo, 2011; Ozubko et al., 2012). This also follows from previous memory literature showing that children can retain information in long-term memory after a 1-week delay (Holland et al., 2015; Scarf et al., 2013). Our second hypothesis was that children are able to create distinct encodings, but these are sensitive to children's less-developed cognitive skills and less likely to maintain the distinct encoding at a delay. Thus, we expected a change in performance for *Say* items at delayed testing, resulting in *Say* items showing equal or worse performance at test than *Listen* items at delayed testing. This follows from previous studies using mnemonic tasks with children suggesting that these do not benefit long-term memory (e.g., Krinsky & Krinsky, 1994) and studies on the Production Effect with children (López Assef et al., 2021; Zamuner et al., 2018) in which developmental stage was identified as a factor mediating the presence of the Production Effect.

Our study also included an exploratory analysis (pre-registered) aimed to investigate whether vocabulary score would predict the Production Effect in our participants. Two of the factors proposed to alter the Production Effect are linguistic and experience-related factors (see Zamuner et al., 2017 for a review). Furthermore, studies have found long-term memory and learning benefits in children with larger vocabularies (Daidone & Darcy, 2021; Gathercole et al., 1997; Munro et al., 2012). Vocabulary size in children has been shown to support phonological awareness (Gorman, 2012) and phonological development (Edwards, et al., 2004), contribute to new word learning (Gathercole et al., 1997), and predict performance in spoken word recognition (Law et al., 2017; Munson, 2001). Additionally, 2- and 3-year-olds' vocabulary has also been linked with their ability to retrieve words at different delays (Munro et al., 2012). Thus, we explored a potential difference in performance related to children's vocabulary sizes. By exploring the effect of vocabulary size, we hoped to gain insights into the potential relation between children's linguistic abilities and the presence and/or strength of the Production Effect. We predicted that higher standardized vocabulary scores would be associated with a more robust Production Effect and thus a higher proportion during recall and higher accuracy during recognition for words produced than words heard. The rationale is that larger vocabularies, and thus more robust lexical representations due to higher numbers of connections in the lexicon, could make the task less cognitively demanding for children, thus allowing them to benefit even more from production.

Method

Participants

Participants were 60 English-speaking children aged 5 years ($n = 30$, 13 males, 17 females) and 6 years ($n = 30$, 20 males, 10 females). This age group was chosen as they have consistently shown the Production Effect with familiar words in previous studies (Icht & Mama, 2015; López Assef et al., 2021). The target sample size of 60 was determined by conducting a power analysis prior to data collection, to obtain 80% power to detect an effect size (*d*) of 0.4 at the standard .05 alpha error probability. The effect size was derived from prior work and calculated doing ANOVAs on results from participants of similar age from the study by López Assef et al. (2021) on the Production Effect. Our power analysis was done in two ways: using PANGEA (v0.2; Westfall, 2016) and mixed models using the SimR package (v1.0.7, Green & MacLeod, 2016). See pre-registration materials for more details. Data collection stopped once we reached 30 participants with usable data for each age group. Participants were recruited through online advertisements on social media, childrenhelpingscience.com (website for online studies), and a participant database. Participants were given the option to enter their name in a draw for a gift card once they completed the experiment. The experiment was conducted online through Zoom. All our participants resided in Canada at the time of testing. Most of our participants lived in the provinces of Ontario and Alberta, Canada. While demographics varied across participants, most were of middle- to high-socioeconomic background. Following previous studies, such as Zamuner et al. (2018) and López Assef et al. (2021), participants were required to have a minimum lifetime average of 70% exposure to English ($M = 89\%$, $SD = 12.8$, range = 70–100), to have learned English from birth, and to have not more than two consecutive years of 30+% exposure to another language as estimated from a language background questionnaire completed by parents. Participants were also required to have normal hearing, normal-to-corrected vision, and no history of language impairments or developmental disabilities as determined by parental questionnaire. Lastly, children had to have standardized vocabulary scores on the Expressive One Word Picture Vocabulary Tests (EOWPVT-4, Martin & Brownell, 2011) that were not lower than 85 (1 standard deviation below the mean). Other than the 60 participants included in the analysis, three additional participants were tested but not included in the analysis for having vocabulary scores below 85. Eleven additional participants were tested but excluded for speaking during all heard trials ($n = 1$), not completing the experiment ($n = 1$), and not participating in both test sessions ($n = 9$).

Stimuli

Test stimuli consisted of 20 monosyllabic English words paired with images (*dog, bed, tree, boat, pig, cow, car, train, duck, frog, truck, chair, shoe, fish, cat, door, horse, book, bee, sock*), with an additional 4 disyllabic words used for practice trials (*apple, dolphin, cookie, flower*). Test words were chosen to have a high percentage (96%) of children that produced the words at 30 months old based on norms available in Wordbank (Frank et al., 2016). 10 words were assigned for each training condition. There were an additional 20 English monosyllabic words (audio only) for the Old/New recognition task. Stimuli were pre-recorded by two female native speakers of English (one speaker recorded items for training, another speaker recorded items for the recognition task) and normalized for amplitude (70 dB). Visual stimuli were colored clipart. Images and sound files can be found in the dedicated OSF repository https://osf.io/gxphj/.

Design

There were two training conditions: *Listen* where participants saw the picture and heard an audio recording of the label of the image and *Say* where participants saw a picture on the screen and name the picture aloud. These were presented in a Block design such that a given child was, for example, presented all the *Say* items in Block 1 and all the *Listen* items in Block 2. Participants were randomly assigned to one of four orders which differed in the order of the training conditions (*Listen, Say*) and in the order of the items within each Block. Orders 1 and 2 had the *Say* training condition in Block 1 and *Listen* in Block 2, whereas order 3 and 4 had the reverse. A single item could appear in the *Say* condition in one Order and in the *Listen* condition in another Order. Tables with information for each order can be found on OSF repository. During the recognition task, participants heard 20 new items (items which were not part of training) and 20 old items from training. Recognition items were not blocked, but pseudorandomized (see OSF for table with descriptions for each list).

Due to the COVID-19 pandemic, this study was conducted on Zoom. Participants were asked to join the experiment in a room without external distractions. Once participants joined the Zoom call and the experimenter checked that the audio and video were working properly, experimenters used the screen share feature to conduct the experiment and vocabulary test using PowerPoint. The experiment was divided into two sessions: Immediate and Delayed testing (1-week delay). Sessions were audio and video recorded for offline coding.

Before training started, participants were told that they would be playing a game. Children saw a clipart image of a boy and were told that they would be helping him put all his toys in a box. Participants were told that there was a challenge in the game: in one part they would hear the child's mom name his toys (*Listen* condition) and in the other part they would have to name the toys that appeared on the screen (*Say* condition). The training portion of the experiment was divided into two Blocks, one for each training condition. There were two practice trials for each training condition, followed by the 10 training trials. If participants made a mistake during practice (for example, producing *Listen* items), these trials would be repeated until the correct response was provided (this was done for 2 participants, both cases for *Listen* practice trials). During the presentation of the items, children saw a picture of a toy box at the bottom of the screen, and then the toy appeared above it, on the center of the screen. Once the child had performed the required action (listening or speaking aloud), the toy was animated to slide into the toy box. Once the training portion was finished (both blocks completed), participants completed a free recall test, where they were asked to name the toys that went inside the box. After each recalled word (regardless of accuracy), a star animation appeared on the screen to keep the child engaged. Script and examples of stimuli can be found on OSF.

Delayed testing was scheduled 1 week after Immediate testing (range = 6–9 days). The session began with a free recall task. Children were told that the boy they helped last week wanted to know if they (the participating child) remembered which toys had gone inside the box the week before. Once the child indicated that they were done and did not remember any more toys, they moved to the Old/New recognition task. This task was added during the second session (compared to López Assef et al. 2021, who only used free recall) to address the possibility that the delay might increase task difficulty, causing children's recall to be at floor level. Previous studies have found different results between free recall tasks and recognition tests (MacLeod & Bodner, 2017; Zamuner et al., 2018). During the Old/New recognition task, children were told that they would hear the boy's sister name some of his toys and that they had to say "Yes" (or indicate with a "thumbs up" gesture) if the toy they heard corresponded to one of the toys that had gone into the box the week before and to say "No" (or do a thumbs down gesture) if it corresponded to a new toy that had not gone inside the box the week before. Participants were rewarded with stars after every five trials. Once the recognition task was over, participants completed the standardized vocabulary test (EOWPVT).

Results

Data and code have been made publicly available on OSF at the following link: https://osf.io/gxphj/.

Coding

Testing was conducted by the first or second authors, who also did a first pass at coding responses. The main coding was done based on audio-video recordings, by a research assistant who was blind to the initial coding done by the two experimenters. Discrepancies between both codings were resolved by an additional research assistant. This additional research assistant only had access to the recordings of trials with discrepancies between the experimenters and main coder; thus, the additional RA was blind to the rest of the trials. There were 10 recalled words and 13 Old/New recognition trials that had discrepancies between the initial two codings. Coding that matched the additional research assistant's coding was marked as correct and resolved. The main coding with resolved discrepancies was the data used for the analyses.

Training trials were coded to ensure that participants completed training properly. Items with training errors were excluded from the analysis ($n = 57$): these were excluded for producing *Listen* items ($n = 13$), elaborating during *Say* items (for example, saying "I like socks," instead of just "socks") ($n = 33$), trials where children were looking away from the screen or loud noises or distractions were happing during the trial ($n = 2$), technical issues ($n = 1$), producing the word during training more than once ($n = 1$), and other errors ($n = 7$). Training errors represented 5% of our total trials. Exclusions from the recall task included practice items from training (*apple, dolphin, cookie, flower*) and extra recalled words (uttered by a child during recall but not part of training). Words that were recalled more than once were counted as a single recall. Some variation was allowed during training and the free recall task, for example, saying *kitty* for *cat*, although the trials corresponding to these words were excluded from the recognition analysis as the word produced during training did not match the auditory stimuli (*cat*) heard during recognition ($n = 2$).

Free recall task

We used a generalized linear model framework to examine whether training condition and test session significantly predicted recall rates. This analysis followed our pre-registration. Our dependent variable was recall (yes, no). Our fixed effects were Training Condition (*Listen, Say*), Test Session (*Immediate, Delayed*), and Training Order (i.e., order of Training Conditions in Blocks: *Listen in Block 2, Say in Block 2*). Training Order was included as a control variable to account for potential recency effects. Results from the recall task are shown in Fig. 1.

All analyses were performed in R (v4.2.0, R Core Team) using the logistic mixed effect model function from the lme4 package (v1.1-27.1, Bates et al., 2015). We started from the most complex model structure with a random structure containing random by-participant intercepts, by-participant slopes for Training Condition and Test Session and their interaction, random by-item intercepts, and by-item slopes for Training Condition, Test Session, Training Order, and their interactions. Models were simplified to account for convergence issues and singularity errors, following our pre-registered analysis, until a model converged. Table 1 provides results from the model, along with the final model structure.

There was a significant main effect of Test Session ($\beta = -0.49$, $SE = 0.11$, $p = <.001$), a significant interaction between Training Condition and Training Order ($\beta = -0.86$, $SE = 0.36$, $p = .016$), and a 3-way interaction between Training Condition, Test Session, and Training Order ($\beta = 1.7$, $SE = 0.45$, $p = <.001$). No main effect of Training Condition was found ($\beta = -0.34$, $SE = 0.19$, $p = .07$). Post hoc comparisons were done on the final model with the emmeans package, using asymptotic estimations for degrees of freedom (v1.7.4-1; Lenth, 2020). Estimated marginal means back-transformed into proportions are reported. Post hoc tests for the significant effect of Test Session revealed a higher probability of recall for words at Immediate Testing ($M = 0.19$, $SE = 0.02$, *95% CI* [0.15–0.24]) than at Delayed Testing ($M = 0.13$, $SE = 0.02$, *95% CI* [0.10–0.16]). The interaction between Training Condition and Training Order (Table 2) showed a significant difference between recall across Training Conditions when *Say* was in Block 2 (*Say* $M = 0.22$, $SE = 0.03$, *95% CI* [0.17–0.27]; *Listen* $M = 0.11$, $SE = 0.02$, *95% CI* [0.07–0.17]), but not when *Listen* was in Block 2 (*Say* $M = 0.15$, $SE = 0.21$, *95% CI* [0.11–0.19]; *Listen* $M = 0.16$, $SE = 0.03$, *95% CI* [0.11–0.23]).

Lastly, post hoc tests for the 3-way interaction between Training Condition, Test Session, and Training Order are plotted in Fig. 2 and presented in Table 3. There was an overall recency effect at Immediate Testing, in which the Training Condition presented in Block 2 showed higher recall. This recency effect was not present at Delayed Testing. At Immediate Testing in Block 1, there was no significant difference in recall proportion for items presented across Training Conditions. Average recall proportion for the Block 1 *Say* training condition was 0.16 ($SD = 0.37$) and for the Block 1 *Listen* training condition was 0.15 ($SD = 0.36$). On the other hand, for items that appeared in Block 2, *Say* items had an average recall proportion of 0.32 ($SD = 0.47$), while the average for *Listen* items in Block 2 was 0.27 ($SD = 0.44$), which was a significant difference. This suggests that *Say* items in Block 2 received a bigger memory boost than *Listen* items in Block 2, showing a Production Effect on top of the recency effect.

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

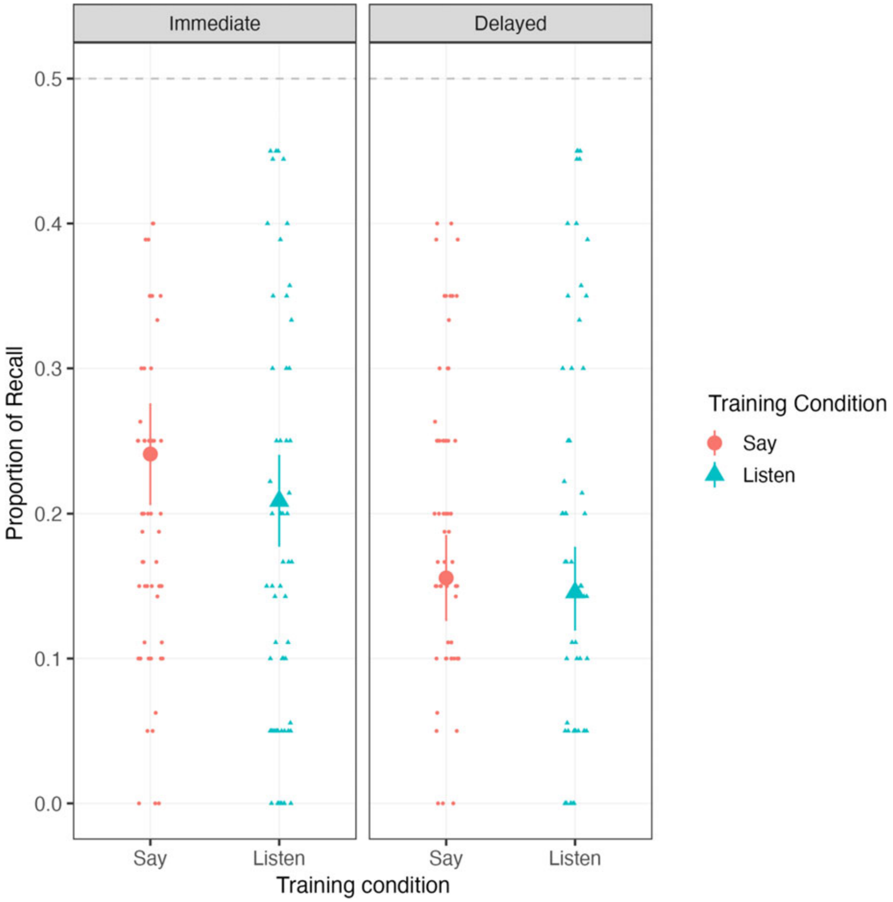

Figure 1. Proportion of recall by training condition (Listen, say) and test session (Immediate, delayed). *Note.* Points are the condition means by participant with error bars indicating 95% confidence intervals. Dotted line indicates 50% chance level.

Old/New recognition task

During the Old/New recognition task, children heard old test words from the week prior, along with new words in a randomized order. They were instructed to answer "Yes" if the word they heard corresponded to one of the toys from the week before and to answer "No" if it was a new word. Following our pre-registration, we excluded a subset of participants ($n = 15$) who had an equal or higher proportion of "Yes" answers for new items over old items (Bernard & Onishi, 2023). These children were excluded because it was not clear how well they understood the recognition task given that they were equally or more likely to answer "Yes" (i.e., "I saw this during training last week") for New items that were not part of training, than for items they actually heard in training. After exclusions, recognition analysis included data from 45 participants. Statistical analysis was restricted to old items only (items from training) since we are interested in the comparison between previously trained *Say* and *Listen* items. The

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

Table 1. Results from linear mixed model estimating free recall by training condition (Listen, say), test session (Immediate, delayed), and training order (Listen in Block 2, say in Block 2)

Fixed effects	Estimate	*SE*	*z* value	*p*-value
Training Condition	0.34	0.18	1.83	.06
Test Session	−0.49	0.11	−4.25	<.001
Training Order	−0.05	0.15	−0.33	.73
TrainingCondition*TestSession	−0.05	0.22	−0.22	.82
TrainingCondition*TrainingOrder	−0.86	0.36	−2.38	<.05
TestSession*TrainingOrder	−0.06	0.22	−0.27	.79
TrainingCondition*TestSession*TrainingOrder	1.7	0.45	3.77	<.001

Note. The final model had the following syntax specified in the lme4 package: recall_yes_na ∼ traincondition_sum*session_sum*trainorder_sum + (traincondition_sum|participant) + (1|target). The proportion of variance accounted for by the final model (pseudo-R^2) was calculated using the r.squaredGLMM function: fixed effects (marginal theoretical $R^2_m = 0.04$); fixed and random effects (conditional theoretical $R^2_c = 0.20$).

Table 2. Post hoc tests of the estimated marginal means for significant 2-way interaction from the model predicting proportion of recall by training condition (Listen, say) and training order (Listen in Block 2, say in Block 2)

	Odds ratio	*SE*	*df*	*z* ratio	*p*-value
Listen in Block 2	0.91	0.24	Inf	−0.38	.70
Say in Block 2	2.18	0.57	Inf	2.97	.003

Note. Results are averaged over the levels of Test Session; tests were performed on the log odds ratio scale. Odds ratio represents a comparison of performance on *Say* to *Listen* training conditions (*Say/Listen*).

recognition analysis followed the same procedure as for our recall task, except for the exclusion of Test Session in the model, because the Old/New recognition task was only done at Delayed Testing. Accuracy was the dependent variable (correct, incorrect). Results from the Old/New recognition task are shown in Fig. 3. Average accuracy for *Say* items was 0.46 ($SD = 0.50$), and for *Listen* items was 0.48 ($SD = 0.50$). Due to results for this task hovering around chance, we used d-prime scores (found on OSF repository) to further analyze participants' performance in this task and to make sure they had completed the task correctly (i.e., identified *Old* items over *New* items). All of our participants had d-prime scores above 0 ($M = 0.98$, $SD = 0.42$), suggesting that their performance is above chance, and thus, their low recognition rates across Training Conditions were not due to task-related issues, since they were able to accurately distinguish between *Old* and *New* items, but because of low memory for items in both Training Conditions. The initial model included Training Condition as a fixed effect, with Training Order as a control variable, along with random by-participant intercepts, by-participant slopes for Training Condition, by-item intercepts, and by-item slopes for Training Condition, Training Order, and the interaction between them. The model was simplified following the same procedure as the previous recall analysis. There were no significant main effects or interactions (Table 4).

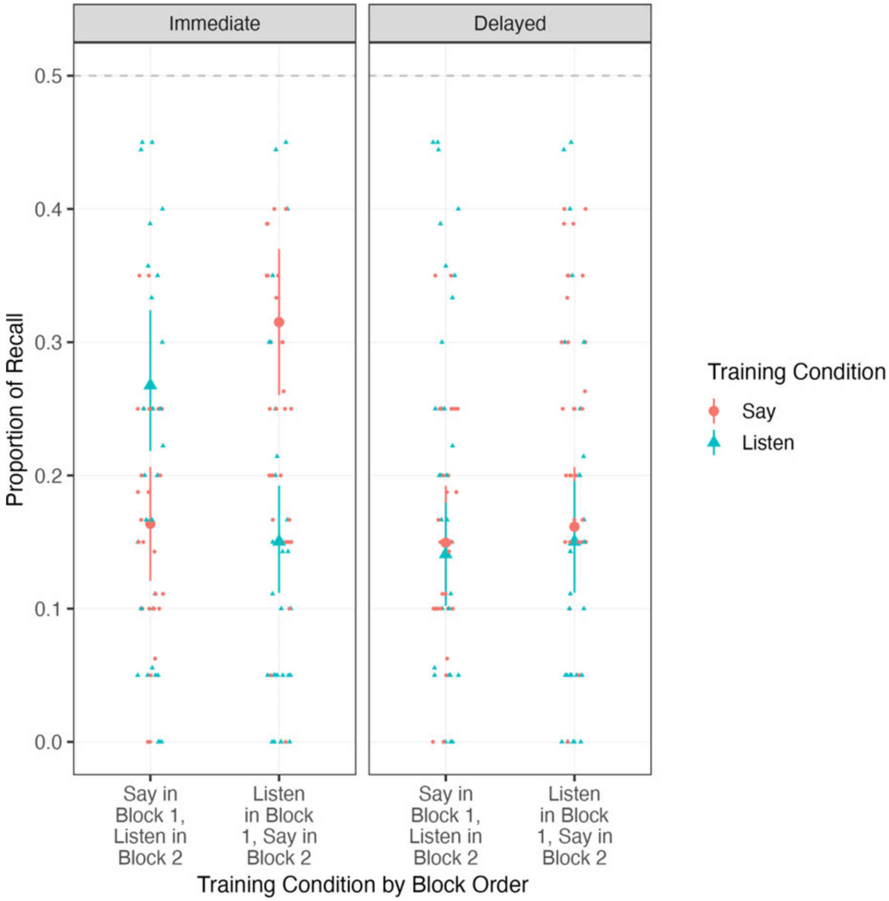

Figure 2. Mean recall by training condition, test session, and training order.
Note. Points are the recall means by participant with error bars indicating 95% confidence intervals. Dotted line indicates 50% chance level.

Exploratory analysis: vocabulary score

Statistical analyses for incorporating Vocabulary Scores followed the same procedure as the recall analysis. The analysis was conducted on recall data from Immediate Testing only, as this was the only instance where we found a Production Effect. Additionally, visualizations of the vocabulary data (found on OSF) suggested similar performance across Vocabulary Scores at Delayed Testing, but possible differences at Immediate Testing. The variable for standardized Vocabulary Score was centered and used as a fixed effect, with proportion of recall as the dependent variable. We used the standardized Vocabulary Score from the EOWPVT as opposed to the raw scores because the standardized score controls for age. The initial model included Training Condition, Training Order, and Vocabulary Score as fixed effects; Training Order as a control variable; random by-participant intercepts, by-participant slopes for Training Condition, by-item intercepts, and

Table 3. Post hoc tests of the estimated marginal means for significant 3-way interaction from the model predicting proportion of recall by training condition (Listen, say), test session (Immediate, delayed), and training order (Listen in Block 2, say in Block 2)

	Odds ratio	*SE*	*df*	*Z* ratio	*p*-value
Immediate Testing					
Listen in Block 2	0.61	0.18	Inf	−1.69	.09
Say in Block 2	3.43	1.03	Inf	4.14	<.0001
Delayed Testing					
Listen in Block 2	1.35	0.43	Inf	0.95	.34
Say in Block 2	1.38	0.43	Inf	1.02	.31
	M	*SE*	*df*	*95% CI*	
Immediate Testing					
Listen in Block 2					
Listen	0.23	0.04	Inf	[0.16–0.32]	
Say	0.15	0.03	Inf	[0.11–0.21]	
Say in Block 2					
Listen	0.11	0.03	Inf	[0.07–0.18]	
Say	0.30	0.04	Inf	[0.24–0.39]	
Delayed Testing					
Listen in Block 2					
Listen	0.11	0.03	Inf	[0.07–0.17]	
Say	0.14	0.03	Inf	[0.10–0.20]	
Say in Block 2					
Listen	0.11	0.03	Inf	[0.07–0.18]	
Say	0.15	0.03	Inf	[0.10–0.21]	

Note. Tests were performed on the log odds ratio scale. Odds ratio represents a comparison of performance on *Say* to *Listen* training conditions (*Say/Listen*).

by-item slopes for Training Condition, Training Order, Vocabulary size, and the interaction between them. Results for our exploratory analysis are shown in Fig. 4. There was a significant interaction between Training Condition and Training Order ($\beta = -1.74$, $SE = 0.40$, $p = <.001$), as seen in the main analyses, but no effects of Vocabulary Score (the corresponding model and results can be found in OSF repository).

As stated in our pre-registration, we also analyzed the effect of vocabulary using a production effect score to see whether vocabulary size affected the strength of the Production Effect. This was calculated from the difference between proportions of recall across both training condition, by participant. A negative score would indicate a child with a Reverse Production Effect (i.e., higher recall for *Listen* than *Say*), whereas a positive score shows the classic Production Effect (i.e., higher recall for

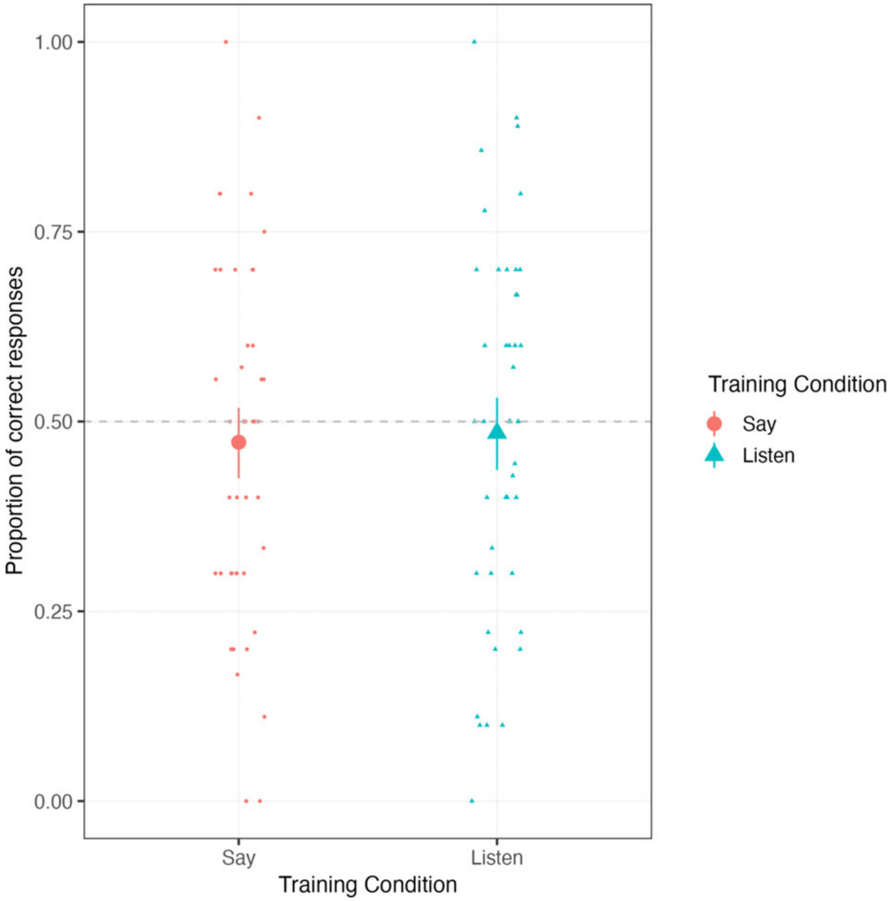

Figure 3. Proportion of correct responses during Old/New recognition task by training condition (Listen, say).
Note. Recall was only tested after 1-week delay. Points with error bars indicating 95% confidence intervals are the condition means. Dots are proportion of correct responses by participant. Dotted line indicates 50% chance level.

Say than for *Listen*). Our analysis was done using the lm() function from the stats package (v4.2.0, R Core Team). The production effect score was our dependent variable, with standardized Vocabulary Score and Training Order as our fixed effects (Training Condition is not present here since each participant only has one score). This model did not include item-level effects because each participant only had one score. Results showed only a significant main effect of Training Order ($\beta = -2.59, SE = 0.61, p = <.001$). When *Listen* was in Block 2, participants were more likely to show a negative Production Effect score, whereas when *Say* was in Block 2, they were more likely to show a positive score. These results pattern with those of our recall task. Importantly for our exploratory analyses and the factors that we were examining, vocabulary and the interaction between Vocabulary Score and Training Order were not significant (full results in OSF repository).

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

Table 4. Results from linear mixed model estimating recognition accuracy by training condition (Listen, say) and training order (Listen in Block 2, say in Block 2)

Fixed Effects	Estimate	*SE*	z value	*p*-value
Training Condition	−0.07	0.17	−0.41	.68
Training Order	0.52	0.29	1.80	.07
TrainingCondition*TrainingOrder	−0.30	0.31	−0.96	.34

Note. The final model had the following syntax specified in the lme4 package: recognition~ traincondition_sum*trainorder_sum + (1|participant) + (traincondition_sum|target). The proportion of variance accounted for by the final model (pseudo-R^2) was calculated using the r.squaredGLMM function: fixed effects (marginal theoretical $R^2{}_m = 0.01$); fixed and random effects (conditional theoretical $R^2{}_c = 0.21$).

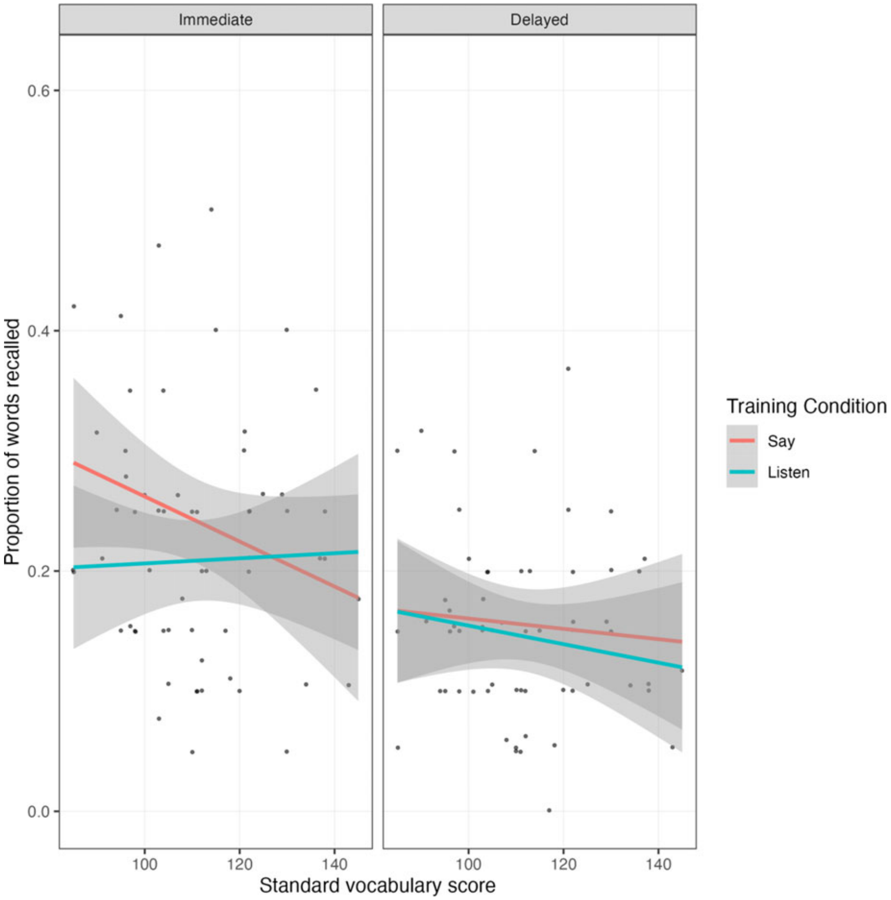

Figure 4. Proportion of recall by training condition (Listen, say) and standardized vocabulary score.
Note. Points are the condition means by participant with error bars indicating 95% confidence intervals.

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

Based on a suggestion from a reviewer, we also explored the effect of amount of exposure to English on the Production Effect, and we also included Age in this exploratory analysis to tease apart any confounding effects between amount of exposure to English and age. Exposure to English was calculated as the lifetime percentage of English (based on hours), as measured in our Language Background Questionnaire. Age was calculated in months. This exploratory analysis is available on OSF. The correlation between Exposure to English and Age was significant ($p = < .001$), however not highly correlated ($r = -.056$). Models that included both Exposure to English and Age did not converge; therefore, we created separate models. We created a simple model with Training Condition and Exposure to English and compared this to another model with Training Condition and Age to explore the effect of each separately. These analyses were based only on the data from the Immediate Testing session. Results showed a significant effect of Exposure to English on recall ($\beta = -0.017, SE = 0.006, p = < .001$), with overall recall decreasing as the percentage of exposure to English increased. There was no significant interaction between Exposure to English and Training Condition. It is not clear why children with less language exposure recalled more words, but interpretation of this effect must be made cautiously, as most participants clustered at the higher end of language exposure (70% to 100%). Only six participants had less than 75% exposure to English, with a high degree of variability on recall rates. When we exclude those six participants, the effect of language exposure on overall recall is no longer significant, but the model shows a significant interaction between Training Condition and Language Exposure ($\beta = -0.73$, $SE = 0.30$, $p = .012$). Based on these results, if we tested children along the full continuum of language exposure, one would expect a negative correlation between the amount of language exposure and overall recall.

The model for Age had a significant interaction between Training Condition and Age ($\beta = -0.07$, $SE = 0.018$, $p = <.0001$), with children between 60 to 72 months showing a recall advantage for *Say* items, whereas for older participants (73 to 83 months) this advantage was not present. This pattern is comparable to a previous study with familiar words using free recall by López Assef et al. (2021). In that study, there was an advantage for *Listen* items for 3 and 4-year-olds (36 to 59 months), which shifted to an advantage for *Say* items for 5- and 6-year-olds. The production advantage shift began at 61 months; however, it was only statistically significant for participants aged 75.4 months and older. Thus, the timing of the appearance of a significant Production Effect is slightly different across the two studies, showing variability across 5- and 6-year-olds which arguably varies depending on the stimuli and the task. While this is an initial exploration, this gives further indication of possible age-related factors on the Production Effect in children. However, as many variables are often related, more research is needed. These studies would need to continue to carefully control for factors such as stimuli and experiment design, moreover to adequately include participants across a range of language exposure, age, vocabulary size, and training order (due to the significant recency effect).

Discussion

In the current study, we investigated whether the Production Effect would remain after a 1-week delay in children aged 5 and 6 years old. Results showed a recency effect at Immediate Testing, in which the Training Condition that appeared in Block 2 showed higher recall. When comparing recalled items at Immediate Testing in Block 2, we found a recall advantage for produced words (*Say* condition) over heard words (*Listen* condition); however, no advantage for any Training Condition was found after a 1-week delay. An Old/New recognition task done after a 1-week delay also showed no Training Condition advantage, with similar recognition accuracy in both training conditions.

Thus, our results are in contrast with previous studies with adults which found the Production Effect held at delayed testing (Grohe & Weber, 2018; Icht & Mama, 2019; Kaushanskaya & Yoo, 2011; Ozubko et al., 2012). Our results suggest saying the items aloud can create distinct encodings for produced words at immediate testing, although the success in creating these encodings can also be affected by recency. However, this memory trace appears to decay over time, possibly due to the cognitive limitations stemming from developmental characteristics of children, resulting in difficulty maintaining or retrieving the distinct encoding from memory. While the lack of the Production Effect at Delayed testing was somewhat surprising, previous studies have shown that for children, memory advantages stemming from a mnemonic task could be restricted to immediate testing and do not seem to aid long-term memory (Krinsky & Krinsky, 1994). Krinsky and Krinsky (1994) investigated the effect of using mnemonic training on children (ages 10 to 12) in two experiments. Children were trained on how to use a mnemonic device and were tested on recall for familiar, high-frequency nouns. Results showed an increase in recall after being trained on how to use the mnemonic, suggesting that children could benefit from using mnemonics. Interestingly, this benefit was restricted to immediate testing and was not found at delayed testing. At delayed testing, children showed similar recall to what they showed before training, suggesting that mnemonic task benefits could be restricted to immediate testing for children.

To date, studies with children paint a complex picture for the Production Effect. While studies support the theory that production can create distinctive encodings, whether this occurs and if it is possible to maintain in time seem to be affected by multiple factors. In the current study, we explored one possible factor that could affect the Production Effect: vocabulary size. Although previous proposals have hypothesized that language-related factors may influence the Production Effect (Zamuner et al., 2017; see Vihman, 2022 for a discussion on the effect of vocal production on word learning in infants and children), the current results fail to support this hypothesis. For our participants, Vocabulary Score was not a predictor of the Production Effect. While we did not find any significant effects for Vocabulary Score, Fig. 4, showing the proportion of recall by Training Condition across different Vocabulary Scores, suggests that participants with lower vocabulary scores benefit from production more than those on the higher end of the scale and thus show a Production Effect. Further research could look at different types of measures for vocabulary and other language skill indicators to explore if any other measure may be a clearer indicator. Furthermore, since we excluded participants who fell below 1

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

standard deviation from the mean Vocabulary Score, it is possible that vocabulary size could predict the Production Effect for children with lower scores.

Another approach to address how language experience might relate to the Production Effect would be to test bilingual and/or second language learners. Comparing monolingual children to other groups would also tap into a different type of experience which is not measured by vocabulary tests. Knowing another language and/or having less experience or proficiency with a language could impact the likelihood of the Production Effect. This may be similar to what has been found in previous studies with adults: the Production Effect has been shown to be impacted by language experience in the form of proficiency (Baese-Berk & Samuel, 2016) and degree of familiarity with a speaker's accent (Grohe & Weber, 2018). Differences in language experience with children could affect the activation of information during the learning process and interact with the Production Effect. For example, Grohe and Weber found that, overall, new words presented in familiar accents were recalled more than new words presented in unfamiliar accents. They argue that for new words presented in familiar accents, adult learners activate accent-related information, and this facilitates processing. This information is not available for words in unfamiliar accents. Applying this to the Production Effect with children, differences in language experience might also lead to differences in how information is activated or processed, possibly causing different effects for production on language learning or memorization. Moreover, L2 learners may not have accurate vocal productions or may not be able to create distinctive representations in their L2 in the same way that monolingual speakers would.

Although it was not part of our original hypotheses, we also found a main effect of recency in our analyses since we included the order of the blocked conditions as a control. We found a significant interaction between Training Condition and Block Order for the recall test after Immediate Testing. When the *Listen* condition was in Block 1 and *Say* in Block 2, more items were recalled from the *Say* condition. Conversely, when the *Say* condition was in Block 1 and *Listen* in Block 2, more items were recalled from the *Listen* condition. However, when looking at recall rates with just Block 1, there were similar recall rates for both training conditions (see Fig. 2). In contrast, for just Block 2, recall rates for the training conditions were statistically different from each other: produced words from the second Block showed a higher recall than heard words from the second Block. Thus, *Say* items assigned to the second Block of the study received a bigger memory boost for the free recall task compared to *Listen* items assigned to the second Block. This interaction between order and production has not been previously reported in the studies with children; however, of the four existing studies with children, most have used a mixed-condition design rather than a block design (mixed-condition design: Icht & Mama, 2015; Pritchard et al., 2019; Experiment 1 in Zamuner et al., 2018; block-condition design: López Assef et al., 2021; Pritchard et al., 2019; Experiment 2 in Zamuner et al., 2018). Focusing on just the studies with a block design which analyzed block order effects, Pritchard et al., (2019) found no main effect of block order using mixed ANOVAs. It is possible that if children create distinctive encodings for produced items in Block 1, these could be more difficult to maintain in memory during the second half of the experiment due to cognitive load effects. Children in this scenario would have to retain distinctive representations for words from Block 1

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

in memory, while also having to attend to the new information from Block 2. This could lead to forgetting some of the information from Block 1 to prioritize completing the task in Block 2. In this case, children would more easily recall *Listen* items from Block 2 (recency effect), over the *Say* items from Block 1. While unexpected, recent studies have also found a recency advantage in Production Effect designs with adults, where the last few items show better recall (Cyr et al., 2021; Gionet et al., 2022; Saint-Aubin et al., 2021). Most relevant to our study, in adults, produced items showed worse recall when they appeared earlier in the study compared to silently studied items (similar to our results) and showed better recall at later positions. This is relevant for future research looking into the Production Effect, as it is possible that reversal effects or disadvantages for produced items in blocked within-subject designs are due to order effects, and not specific to speech production. However, we did still find within Block 2, a significant Production Effect in the predicted direction. More research on the role of order effects and blocked vs. mixed designs is needed to get a better understanding of their interaction with the Production Effect. Additionally, this opens the question of other types of non-linguistic factors that interact with the Production Effect.

One thing to consider is that our experiment was conducted online, instead of in-person, like previous child and adult studies, which could have caused participants to be less engaged or more distracted during the tasks. We addressed this concern by framing our experiment as a game for children and creating a storyline that would help them become more interested in the task. Our participants showed overall a higher percentage of recall for both training conditions at Immediate testing (24% for *Say* words and 20% for *Listen* words), than the 5- and 6-year-olds in the in-person López Assef et al. (2021) study (18% for *Say* items, 13% for *Listen* items), suggesting that the switch from in-person testing to online testing did not render the task more difficult. Our results are also in line with other studies on the Production Effect with similar-aged children: Icht & Mama's (2019) participants (5-year-olds) recalled 29% of *Say* words and 21% of *Listen* items in Experiment 1 and recognized 54.2% of *Say* items and 40.4% of *Listen* items in Experiment 2. Participants in Zamuner et al. (2018), aged 4.5 to 6 years old, recalled an average of 1.7 out of 4 words (42.5%) for *Say* items and 0.21 out of 4 words (5.25%) for *Listen* items.

While immediate recall was not affected by the change to online testing, it is possible that because there were more distractions during our experiment, from children not being in a controlled environment, that were impediments to creating long-lasting distinctive encodings. Furthermore, the location in which encoding took place could affect recall. We did not control continuity in the environment where Immediate and Delayed testing took place (e.g., changing from the living room to an office between sessions, although this was not common). Replicating the experiment and having participants come into a lab twice might provide insights into whether continuity in the training and testing environment affected encoding and/or recall.

Furthermore, while our recall task was below chance (although showing similar rates to previous studies), participants were above chance in our recognition task. It is possible that free recall, retrieving a word from memory with no additional aid, might be too difficult of a task for children and thus not be fully appropriate to investigate the Production Effect. Other tasks, such as our recognition task, might be easier for children and thus allow us to tap into the Production Effect without large task-related confounding effects.

To conclude, our findings contrast with adult studies which showed the Production Effect after a delay (Grohe & Weber, 2018; Icht & Mama, 2019; Kaushanskaya & Yoo, 2022; Ozubko et al., 2012) and add to the growing literature on the Production Effect in younger participants. So far studies with children have shown that the Production Effect arises with familiar words (Icht & Mama, 2015; older participants in López Assef et al., 2021) and written stimuli with elementary school-age children (Pritchard et al., 2019), but a Reversed Production Effect is found with pre-school children and novel words (Zamuner et al., 2018) and for known words for 2- and 3-year-olds (López Assef et al., 2021). The general pattern seen in child studies is that when cognitive demands are high, whether this is due to developmental, language, cognitive issues (or a combination of them), the memory for produced items cannot be predicted solely by relying on distinctiveness (otherwise child and adult studies would consistently show a Production Effect). Our results support the hypothesis that children can create distinctive encodings for produced items, while also adding that it is possible that these encodings are not robust enough to last over long delays. This could be caused by either issues during encoding or at later stages (e.g., when transferring information from short-term memory to long-term memory). Additionally, our design only included one exposure to each training item, and it is possible that increasing the number of repetitions could provide children with more information or opportunities to create a robust memory trace. Future research looking at cognitive skills and additional measures is needed to better understand the Production Effect in both children and adults. Work on the Production Effect not only leads to a better understanding of the role of speech production on language learning but also provides further insight into the type of information that is encoded in lexical representations or in memory and that can later be used to retrieve items. This is especially relevant for psycholinguistic models of lexical representations, since if language experience or production can be activated and influence retrieval, then it is possible they are part of a word's representation. Additionally, the Production Effect has practical uses in real-life situations, as it can be used for pedagogical purposes as a mnemonic device.

Replication package. Data and code have been made publicly available on OSF at the following link: https://osf.io/gxphj/.

Acknowledgements. This research was supported by Agencia Nacional de Investigación y Desarrollo (ANID, Chilean Government) under grant 72200366 awarded to Belén López Assef. The authors thank our participants, Rachel Miller and members of the uOttawa Centre for Child Language Research who contributed to this research.

Competing interests. Research was supported by a Doctoral Fellowship of the Chilean Agencia Nacional de Investigación y Desarrollo (#72200366) granted to Belén López Assef. Other authors declare no interests.

References

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, **81**(4), 981–1005. https://doi.org/10.3758/s13414-019-01725-4

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, **89**, 23–36. https://doi.org/10.1016/j.jml.2015.10.008

Baese-Berk, M. M., & Samuel, A. G. (2022). Just give it time: Differential effects of disruption and delay on perceptual learning. *Attention, Perception, & Psychophysics*, **84**(3), 960–980. https://doi.org/10.3758/s13414-022-02463-w

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

Bemis, R. H., & Leichtman, M. D. (2019). That was last time! The effect of a delay on children's episodic memories of learning new facts. *Infant and Child Development*, **28**(1), e2113–n/a. https://doi.org/10.1002/icd.2113

Bernard, A., & Onishi, K.H. (2023). Novel phonotactic learning by children and infants: Generalizing syllable-position but not co-occurrence regularities. *Journal of Experimental Child Psychology*. https://doi.org/10.1016/j.jecp.2022.105493

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, **38**(6), 1711–1719. https://doi.org/10.1037/a0028466

Cho, K. W., & Feldman, L. B. (2016). When repeating aloud enhances episodic memory for spoken words: interactions between production- and perception-derived variability. *Journal of Cognitive Psychology*, **28**(6), 673–683. https://doi.org/10.1080/20445911.2016.1182173

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, **26**, 341–361. https://doi.org/10.1016/0749-596X(87)90118-5

Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2021). The production effect over the long term: Modeling distinctiveness using serial positions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **48**(12), 1797–1820. https://psycnet.apa.org/doi/10.1037/xlm0001093

Daidone, D., & Darcy, I. (2021). Vocabulary size is a key factor in predicting second-language lexical encoding accuracy. *Frontiers in Psychology*, **12**, 2769. https://doi.org/10.3389/fpsyg.2021.688356

Edwards, J, Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, **47**(2), 421–436. https://doi.org/10.1044/1092-4388(2004/034)

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, **44**(3), 677–694. https://doi.org/10.1017/S0305000916000209

Forrin, N. D., & MacLeod, C. M. (2016). Auditory presentation at test does not diminish the production effect in recognition. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale* **70**(2), 116–124. https://doi.org/10.1037/cep0000092

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, **40**(7), 1046–1055. https://doi.org/10.3758/s13421-012-0210-8

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, **16**, 110–119. https://doi.org/10.3758/BF03213478

Gathercole, S.E, Hitch, G. J., Service, E., & Martin, A. J. (1997). Phonological short-term memory and new word learning in children. *Developmental Psychology*, **33**(6), 966–979. https://doi.org/10.1037/0012-1649.33.6.966

Gionet, S., Guitard, D., & Saint-Aubin, J. (2022). The production effect interacts with serial positions: Further evidence from a between-subjects manipulation. *Experimental Psychology*, **69**(1), 12–22. https://psycnet.apa.org/doi/10.1027/1618-3169/a000540

Gorman, B. K. (2012). Relationships between vocabulary size, working memory, and phonological awareness in Spanish-speaking English language learners. *American Journal of Speech-Language Pathology*, **21**(2), 109–123.https://doi.org/10.1044/1058-0360(2011/10-0063)

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, **7**(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Grohe, A. K., & Weber, A. (2018). Memory advantage for produced words and familiar native accents. *Journal of Cognitive Psychology*, **30**(5–6), 570–587. https://doi.org/10.1080/20445911.2018.1499659

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

Holland, A., Simpson, A., and Riggs, K. J. (2015). Young children retain fast mapped object labels better than shape, color, and texture words. *Journal of Experimental Child Psychology*, **134**, 1–11. https://doi.org/10.1016/j.jecp.2015.01.014

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, **11**(4), 534–537. https://doi.org/10.1016/S0022-5371(72)80036-7

Howe, M. L. (2006). Distinctiveness effects in children's memory. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 237–257). New York, NY: Oxford University Press.

Howe, M. L., Courage, M. L., Vernescu, R., & Hunt, M. (2000). Distinctiveness effects in children's long-term retention. *Developmental Psychology*, **36**(6), 778–792. https://psycnet.apa.org/doi/10.1037/0012-1649.36.6.778

Icht, M., Ben-David, N., & Mama, Y. (2020). Using vocal production to improve long-term verbal memory in adults with intellectual disability. *Behavior Modification*, **45**(6), 715–739. https://doi.org/10.1177/0145445520906583

Icht, M., & Mama, Y. (2015). The Production Effect in memory: a prominent mnemonic in children. *Journal of Child Language*, **42**(5), 1102–1124. https://doi.org/10.1017/S0305000914000713

Icht, M., & Mama, Y. (2019). The effect of vocal production on vocabulary learning in a second language. *Language Teaching Research*, **26**(1), 79–98. https://doi.org/10.1177/1362168819883894

Jones, A. C., & Pyc, M. A. (2014). The production effect: Costs and benefits in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **40**(1), 300–305. https://doi.org/10.1037/a0033337

Kan, P. F. (2014). Novel word retention in sequential bilingual children. *Journal of Child Language*, **41**(2), 416–438. https://doi.org/10.1017/S0305000912000761

Kapnoula, E. C., & Samuel, A. G. (2022). Reconciling the contradictory effects of production on word learning: Production may help at first, but it hurts later. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **48**(3), 394–415. https://doi.org/10.1037/xlm0001129

Kapnoula, E.C., & Samuel, A.G. (2023). Wait long and prosper! Delaying production alleviates its detrimental effect on word learning. *Language, Cognition, and Neuroscience*, **38**, 724–744. https://doi.org/10.1080/23273798.2022.2144917

Kaushanskaya, M., & Yoo, J. (2011). Rehearsal effects in adult word learning. *Language and Cognitive Processes*, **26**(1), 121–148. https://doi.org/10.1080/01690965.2010.486579

Krinsky, R., & Krinsky, S. G. (1994). The peg-word mnemonic facilitates immediate but not long-term memory in fifth-grade children. *Contemporary Educational Psychology*, **19**(2), 217–229. https://doi.org/10.1006/ceps.1994.1018

Law, F., Mahr, T., Schneeberg, A., & Edwards, J. (2017). Vocabulary size and auditory word recognition in preschool children. *Applied Psycholinguistics*, **38**(1), 89–125. https://doi.org/10.1017/S0142716416000126

Lawson, M., & London, K. (2015). Tell me everything you discussed: Children's memory for dyadic conversations after a 1-week or a 3-week delay. *Behavioral Sciences & the Law*, **33**(4), 429–445. https://doi.org/10.1002/bsl.2184

Lenth, R. (2020). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.4.4. Retrieved from https://CRAN.R-project.org/package = emmeans

López Assef, B., Desmeules-Trudel, F., Bernard, A., & Zamuner, T. S. (2021). A shift in the direction of the production effect in children aged 2–6 years. *Child Development*, **92**(6), 2447–2464. https://doi.org/10.1111/cdev.13618

López Assef, B., Strahm, S., Boyce, K., Page, M. & Zamuner, T., (2023) Production benefits recall of novel words with frequent, but not infrequent sound patterns. *Glossa: A Journal of General Linguistics*, **8**(1). doi: https://doi.org/10.16995/glossa.8582

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, **26**(4), 390–395. https://doi.org/10.1177/0963721417691356

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **36**(3), 671–685. https://doi.org/10.1037/a0018785

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature*, **385**(6619), 813–815. https://doi.org/10.1038/385813a0

Martin, N. A., & Brownell, R. (2011). *Expressive one-word picture vocabulary test* (4th ed.). Academic Therapy Publications Assessments.

https://doi.org/10.1017/S0142716424000109 Published online by Cambridge University Press

Mayer, K. M., Yildiz, I. B., Macedonia, M., & von Kriegstein, K. (2015). Visual and motor cortices differentially support the translation of foreign language words. *Current Biology*, **25**(4), 530–535. https://doi.org/10.1016/j.cub.2014.11.068

Munro, N., Baker, E., McGregor, K., Docking, K., & Arciuli, J. (2012). Why word learning is not fast. *Frontiers in Psychology*, **3**, 41. https://doi.org/10.3389/fpsyg.2012.00041

Munson, B. (2001). Relationships between vocabulary size and spoken word recognition in children aged 3 to 7. *Contemporary Issues in Communication Science and Disorders*, **28**, 20–29. https://doi.org/10.1044/cicsd_28_s_20

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory*, **20**(7), 717–727. https://doi.org/10.1080/09658211.2012.699070

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **36**(6), 1543–1547. https://doi.org/10.1037/a0020604

Pritchard, V. E., Heron-Delaney, M., Malone, S. A., & MacLeod, C. M. (2019). The production effect improves memory in 7-to 10-year-old children. *Child Development*, **91**(3), 901–913. https://doi.org/10.1111/cdev.13247

Saint-Aubin, J., Yearsley, J. M., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language*, **118**, 104219. https://doi.org/10.1016/j.jml.2021.104219

Sakhon, S, Edwards, K., Luongo, A. A., Murphy, M. M., & Edgin, J. J. (2018). Small sets of novel words are fully retained after 1-week in children with and without Down syndrome: A fast mapping study. *Journal of the International Neuropsychological Society*, **24**(9), 955–965. https://doi.org/10.1017/S1355617718000450

Scarf, D, Gross, J., Colombo, M., & Hayne, H. (2013). To have and to hold: Episodic memory in 3- and 4-year-old children. *Developmental Psychobiology*, **55**(2), 125–132. https://doi.org/10.1002/dev.21004

Vihman, M. M. (2022). The developmental origins of phonological memory. *Psychological Review*, **129**(6), 1495–1508. https://doi.org/10.1037/rev0000354

Vlach, H., & Sandhofer, C. M. (2012). Fast mapping across time: Memory processes support children's retention of learned words. *Frontiers in Psychology*, **3**, 46. https://doi.org/10.3389/fpsyg.2012.00046

Wang, Q., Bui, V.-K., & Song, Q. (2015). Narrative organisation at encoding facilitated children's long-term episodic memory. *Memory*, **23**(4), 602–611. https://doi.org/10.1080/09658211.2014.914229

Wang, J. Y., Weber, F. D., Zinke, K., Inostroza, M., & Born, J. (2018). More effective consolidation of episodic long-term memory in children than adults—unrelated to sleep. *Child Development*, **89**(5), 1720–1734. https://doi.org/10.1111/cdev.12839

Westfall, J. (2016). PANGEA (v0.2): Power analysis for general ANOVA designs [Unpublished manuscript]. Retrieved from http://jakewestfall.org/publications/pangea.pdf

Zamuner, T. S., Morin-Lessard, E., Strahm, S., & Page, M. P. A. (2016). Spoken word recognition of novel words, either produced or only heard during learning. *Journal of Memory and Language*, **89**, 55–67. https://doi.org/10.1016/j.jml.2015.10.003

Zamuner, T. S., Strahm, S., Morin-Lessard, E., & Page, M. P. (2018). Reverse production effect: Children recognize novel words better when they are heard rather than produced. *Developmental Science*, **21**(4), e12636. https://doi.org/10.1111/desc.12636

Zamuner, T. S., Yeung, H. H., & Ducos, M. (2017). The many facets of speech production and its complex effects on phonological processing. *British Journal of Psychology*, **108**(1), 37–39. https://doi.org/10.1111/bjop.12220

Cite this article: López Assef, B., McDonald, M., Bernard, A., & Zamuner, T. S. (2024). Effects of speech production training on memory across short and long delays in 5- and 6-year-olds: A pre-registered study. *Applied Psycholinguistics*. https://doi.org/10.1017/S0142716424000109