This article was accepted by *Child Development* and was published in a revised form at the following link: https://doi.org/10.1111/cdev.13618. This article may be used for non-commercial purposes in accordance with the Wiley Self-Archiving Policy [http://www.wileyauthors.com/self- archiving]

A shift in the direction of the production effect in children aged 2–6 years

Belén López Assef[1], Félix Desmeules-Trudel[2,3], Amélie Bernard[1] and Tania S. Zamuner[1]

[1]Department of Linguistics, University of Ottawa
[2]Department of Psychology, Brain and Mind Institute, The University of Western Ontario
[3] Department of Psychology, University of Toronto, Mississauga

Address for Correspondence:
Belén López Assef
Department of Linguistics
University of Ottawa
Hamelin Hall, 70 Laurier Ave. East
Ottawa ON, Canada K1N 6N5
mblopez075@uottawa.ca

Author Notes
This research was supported by a SSHRC grant awarded to TSZ. BLA was supported by a Doctoral Fellowship of the Chilean Agencia Nacional de Investigación y Desarrollo. AB was supported by Post-doctoral Funding of the Fonds de Recherche du Québec - Société et Culture. This research was conducted at the uOttawa Living Lab at Canada Science and Technology Museum. We thank our participants and Keara Boyce, Myriam Ducos, Leah Gosselin, Tina Ivanov, Emilie Piché, Marianne Roussel and members of the uOttawa Centre for Child Language Research who contributed to this research. We also thank Michal Icht and Yaniv Mama for assistance in replicating aspects of their experimental design. Materials, analysis script and data are available at https://osf.io/8rtxu/?view_only=10c727be2f8f40ca8343d8408421705f

A shift in the direction of the production effect in children aged 2–6 years

When a child learns a word, does it help them encode or remember the word if they say the word aloud? Many of us will answer, 'Yes, of course!' based on our own intuition of the benefit of producing words. But the picture is far from simple during early language development, and the effect of production on language learning is not as straightforward as we might assume. The influence of production starts in early infancy: production-related actions have been found to influence infants' perception of sound contrasts (Bruderer, Danielson, Kandhadai & Werker, 2015; Yeung & Werker, 2013). With children and adults, speech production is both beneficial and detrimental to learning, and the direction of the effect has been argued to depend on task-, attentional-, linguistic- and experience-related factors (Zamuner, Yeung, & Ducos, 2017). The impact of production on word recognition and recall may vary across development, with different effects of production for a child who is just learning how to speak, compared to an older child who is already proficient in producing language. Although production skills begin in infancy and continue to develop throughout childhood, no work on the production effect has been done with children under 4½–5 years of age, and no studies have compared the impact of production across different ages. To further our understanding of the production effect, the current paper investigates the production effect with children across a wide age range, i.e., children aged 2–6 years. Children were trained on images of familiar words in different conditions: LOOK (children see the image in silence), HEARD (children see the image and hear the corresponding word) and PRODUCED (children see the image and produce the corresponding word). Based on prior work, we predicted that older children would recall more words from the PRODUCED than from the LOOK or HEARD training conditions. However, we also expected that if the production effect varies as a function of age, that younger children would

either show an attenuation or reversal of the production effect, extending some of the previous results with older children, and informing us on the specifics of the age-dependent shift in the pattern.

The production effect describes the finding that adults have better recognition and recall for words that are overtly produced compared to words that are read silently (Hopkins & Edwards, 1972; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Multiple studies have replicated the facilitative effect of production with adults in various conditions. For example, the production effect is found when adults mouth, spell, write or type words compared to words read silently; however, it is not found for words that are paired with an overt button press response compared to silent reading (Forrin, MacLeod, & Ozubko, 2012; MacLeod et al., 2010). While the production effect is strongest when adults generate the words themselves, it is also reported when an experimenter or another participant generates the word (MacLeod, 2011). The production effect is seen under different training conditions such as repeating or recalling novel words (Kan, Sadagopan, Janich, & Andrade, 2014; Krishnan, Watkins, & Bishop, 2017), and it also persists after delayed testing (Icht & Mama, 2019; Kaushanskaya & Yoo, 2011). While the production effect has been accounted for in various ways, the original characterization appealed to the concept of distinctiveness - the act of producing items is encoded in memory which then aids later recall (MacLeod et al., 2010; Ozubko, Major, & MacLeod, 2014).

Building on this, Icht and Mama (2015), were the first to investigate the production effect in children. In line with the distinctiveness account, they hypothesized that children's performance on word recognition and recall tasks would be related to the number of unique ways in which words were encoded during training: more distinct encodings (i.e., seen and/or heard and/or produced) would result in greater memory enhancement for those words because there

would be more ways in which those words could be retrieved from memory. In their first experiment, 5-year-old Hebrew-speaking children were trained on words with additive encoding processes: LOOK (visual = see the image), HEARD (visual + audition = hearing the word produced by the experimenter) and PRODUCED (visual + audition + articulation = the execution of a motor action). Note that Icht and Mama used the terms 'look and listen' for HEARD and 'look and say' for PRODUCED, but we use the terms HEARD and PRODUCED to standardize the labelling schemes across the previous and current research. Each participant was presented with thirty images depicted on cue-cards, divided across three boxes that represented each learning action/category (LOOK, HEARD, PRODUCED). On each trial, the experimenter randomly chose a cue-card from one of the three boxes, e.g., a cue-card from the LOOK condition was presented in silence. After training, children had a three-minute break, followed by a free recall task. Recall rates where highest for words that were PRODUCED during training, followed by words that were HEARD, followed by silent observation - LOOK. In a second experiment, they extended this finding to unfamiliar words in two conditions: PRODUCED > HEARD. Thus, the documented production effect in adults (MacLeod et al., 2010), was extended to children (Icht & Mama, 2015). The findings from children were argued to support the distinctiveness account: the memory benefit for produced words was caused by increased distinctiveness (more encoding) compared to words studied under other actions. The greater number of encoding processes resulted in better memory, because children had more information to use during retrieval.

The beneficial effect of production, however, is not always straightforward. In two experiments with same-aged children (4½ –5 years), Zamuner, Strahm, Morin-Lessard, and Page (2018) report on a reverse production effect. Children were trained on auditory novel words paired with visual nonce animals, and asked to either silently listen to the novel words (HEARD)

or to repeat the novel words (PRODUCED). Training was followed by a recognition test using the preferential looking paradigm. Contrary to Icht and Mama (2015), children showed better recognition for the novel words that were HEARD rather than PRODUCED during training. This was unexpected under the distinctiveness account because novel words from the HEARD condition had fewer encoding processes than those from the PRODUCED condition. The finding was also unexpected given that when adults completed the same experiment (Zamuner, Morin-Lessard, Strahm, & Page, 2016), the production effect was observed, with better recognition for the novel words that were PRODUCED versus HEARD during training. At the same time, other work with adults has documented that production does not always benefit recognition and recall (Leach & Samuel, 2007). The attenuation or reversal of the production effect has been found, though not limited to, when using stimuli with unfamiliar sounds (Baese-Berk & Samuel, 2016; Kaushanskaya & Yoo, 2011), unfamiliar accents (Cho & Feldman, 2016, though see Grohe & Weber, 2018), and when varying the testing methodology (Bodner & Taikh, 2012).

Thus, production can be both beneficial and detrimental, and the direction of the production effect depends on various factors, such as the task, attentional resources, linguistic stimuli, and experience-related factors such as age, articulatory skills, and the linguistic background of the participants (Zamuner et al., 2017). Zamuner et al. (2018) hypothesized that the observed reverse-production effect with children aged 4½ –5 years stemmed from the cognitive demands of the learning task which are exacerbated due to the early stages of language development: engaging the production system made it difficult to learn the mapping between the novel word form and referent (also see Munro, Baker, McGregor, Docking, & Arciuli, 2012, for a discussion on cognitive resources during fast mapping). In the HEARD condition, the production system was not engaged, allowing children to direct more attention and processing resources to

mapping, which created stronger representations and improved recognition and recall. Previous researchers have also noted the link between cognitive demands and the influence of production, as stated in the Articulatory Filter Hypothesis: sounds previously produced by learners are more salient in the learners' input than sounds they have not produced (Vihman, DePaolis, & Keren-Portnoy, 2014; Vihman, 2017). The effect of the Articulatory Filter is argued to depend on various factors, such as attentional resources, the processing demands of the task, and the developmental stage of the learner. Processing demands are high when young children are recalling newly learned words; thus, advantages will be seen for words that contain sounds in children's existing production repertoires (also see Richtsmeier & Moore, in press; Zamuner & Thiessen, 2018). However, once learners have mastered production patterns and/or are performing a less demanding task such as passive listening, learners have more processing resources to allocate attention to patterns outside of their production repertoires.

The majority of adult work on the production effect has used reading paradigms; whereas, the two child-based studies (Icht & Mama, 2015; Zamuner et al., 2018) used images and auditory stimuli because their participants were too young to read. Pritchard, Heron-Delaney, Malone, and MacLeod (2019) addressed whether the production effect holds in a reading paradigm with children aged 7–10 years. Children were trained to read words printed in one colour aloud and in the other colour silently (40 in total, 20 from each condition). Children then saw words printed in black (40 trained, 20 untrained) and were asked to verbally indicate if they remembered the words from training. Across two experiments with word and non-word stimuli (words with no visual or semantic referents), Pritchard et al. (2019) found a recognition advantage for items that were read aloud over read silently. This is consistent with the adult literature using reading paradigms; moreover, Pritchard et al. extended the previous findings

from children using images and auditory stimuli (Icht & Mama, 2015) to a reading task. Compared to the results from Zamuner et al. (2018), who found a reverse production effect, Pritchard et al. propose that the difference between their results could also be explained by the cognitive demands for the tasks used across the studies. They suggest that their reading-task and a yes/no recognition task was less cognitively demanding, whereas in Zamuner et al., children had to learn new word forms, map novel words to referent nonce animals, simultaneously produce words, then recognize the nonce animals that corresponded to the novel words. Pritchard et al.'s results are in-line with a distinctiveness approach for produced items, in which production resulted in enhanced learning for school-aged children.

In sum, research on the production effect with children shows mixed results. While studies report a production advantage (Icht & Mama, 2015; Pritchard et al. 2019), others have found the reverse (Zamuner et al., 2018). These differences stem from various interconnected factors: task, attentional resources, linguistic stimuli, and experience. For example, task demands can change across development, as children's language and cognitive skills improve with age. Yet, to date, no studies have taken a developmental approach to investigate the production effect comparing performance at different ages. Here we report on a study that helps fill this gap. We were interested in establishing whether within the same task, the production advantage is constant or changes across development. If children show the same pattern of results across our age range of 2–6 years, then we would show continuity on the effect of production in younger and older children. However, if the direction of the production effect changes (showing attenuation or reversal) at the younger age range, then we would show developmental discontinuity. Furthermore, if we observe a shift in the direction of the production effect within the same general task, this would provide further support for the cognitive-demand hypothesis

which, combined with the distinctiveness approach, gives a more complete picture of the production effect. Experiment 1 was conducted with children aged 3–6 years, and Experiment 2 with 2-year-old children.

Experiment 1

The current experiment investigates the production effect across development; more specifically in children aged between 3–6 years. We adapted Experiment 1 from Icht and Mama (2015). They used a mixed-training design conducted with Hebrew-speaking children aged 5 years, whereas we used a blocked-training design, to have a task more suitable to a wider age range, with English-speaking children aged 3–6 years. Children were trained on familiar words in blocks of three different condition: LOOK (children see the image in silence), HEARD (children see the image and hear the corresponding word) and PRODUCED (children see the image and produce the corresponding word). After a short break, children were tested on a free-recall task, and analyses compared the number of words recalled from each training condition. We predicted that like Icht and Mama, older children would recall more PRODUCED words, followed by HEARD words, followed by LOOK words. If younger children showed a similar pattern, this would indicate a consistent effect of production across development, in a task with familiar words. However, because the same task may be more cognitively demanding for younger children, we could instead find an attenuation or reversal of the production effect for the younger children (more HEARD words recalled, followed by PRODUCED words). If a reverse production effect is observed with our youngest participants, there are different possibilities on how children might recall words from the LOOK condition. LOOK words may show the lowest recall, as silent looking involves the fewest encoding processes according to the distinctiveness account (HEARD > PRODUCED > LOOK). Alternatively, LOOK words may show the highest recall, as silent looking

may be the least cognitively demanding task for younger children (LOOK > HEARD > PRODUCED). Lastly, LOOK words may require more cognitive resources for younger children: when an image is provided without an auditory label, children independently retrieve the word's phonological representation (Ngon & Peperkamp, 2016); this may be more demanding for young children compared to the HEARD condition where the auditory label is provided with the image (HEARD > LOOK). Silent observation and retrieval of a words' phonological representation may also create a less robust memory encoding as compared to retrieving a word and producing it (PRODUCED > LOOK).

Method

Participants

Participants were 120 English-speaking children aged 3–6 years. Age was measured continuously in months, but the breakdown by years was as follows: 3-year-olds *(n* = 30, 16 males, 14 females, *M* = 42 months, *SD* = 3.3, *range* 36–47); 4-year-olds (*n* = 30, 15 males, 15 females, *M* = 53 months, *SD* = 3.7, *range* 48–59); 5-year-olds (*n* = 30, 16 males, 4 females, *M* = 66 months, *SD* = 3.4, *range* 60–71); 6-year-olds (*n* = 30, 8 males, 12 females, *M* = 78 months, *SD* = 3.3, *range* 72–83). Children were tested in a sound-attenuated room at a campus-based lab (*n* = 11) or a museum-based lab (*n* = 109). Participants were recruited from web-based ads or from the museum floor, and received a sticker for participating. Participants were required to have a minimum lifetime average of 70% exposure to English (*M* = 91%, *SD* = 9.1 *, range* 70–100), learned English from birth, and with no more than two consecutive years of +30% exposure to another language as estimated from parental reports. All children were also reported to have normal hearing, normal vision, and no history of language impairment. Twenty-eight additional children were tested but not included in the analyses for the following reasons:

equipment error (n = 2), could not complete the task properly, e.g., produced all words in the LOOK condition (n = 10), no video for off-line coding (n = 3), no recalls (n = 8), did not complete the experiment (n = 3), and had errors on training trials that resulted in the lack of data in a condition (n = 2). Our original design included 2-year-old participants. Data collection with 2-year-olds started (n = 12), but was stopped because of high attrition rates: only one 2-year-old (aged 35 months) completed the task successfully and recalled words. Thus, Experiment 1 reports on data from children aged 3–6 years, and we return to 2-year-olds in Experiment 2.

Stimuli

The test stimuli consisted of 30 monosyllabic English words, which were divided into 3 sets of 10: Set 1 - *ball, boat, cat, cheese, duck, hair, horse, key, moon, train;* Set 2 - *bath, bear, book, chair, cow, fish, pants, shoes, spoon, truck;* Set 3 - *bed, bird, car, door, frog, hat, milk, pig, sock, tree.* There were 6 practice words: *cookie*, *flower*, *apple*, *monkey*, *airplane*, *baby*. All words were controlled for familiarity based on lexical norms from parental reports from the MacArthur-Bates CDI (Fenson et al., 1993; Frank, Braginsky, Yurovsky, & Marchman, 2017). The experimental stimuli are produced on average by 83% (*range* 67–96) of children at 24 months, and 97% (*range* 85–99) of children by 30 months. All stimuli were pre-recorded by a native speaker of English and normalized for amplitude (70dB). The visual stimuli consisted of coloured clipart. To engage the children, we used a stuffed dog (Mr. Wiggles) who wore a removable sleep mask. Twelve WEDGIT™ building blocks were used for the three-minute break between training and the recall task.

Design

There were three training conditions (LOOK, HEARD, PRODUCED), and each training condition was presented in a separate block. Each block consisted of 2 practice trials and 10

training trials, for a total of 36 trials (3 x 12 trials). The order of the block was counterbalanced across participant by using six different lists (each presenting the block conditions in a different order).

The experiment was presented using PowerPoint. At the beginning of each block, an image was displayed on the screen which indicated the task for the upcoming condition: an image of eyes for the LOOK condition, an image of an ear for the HEARD condition, and an image of an open mouth for the PRODUCED condition. This was followed by a display of 'Let's Practice!', the 2 practice trials, then a display of 'Let's go!', followed by the 10 training trials. The practice trials and training trials had the same structure. On each trial, a blank screen was displayed, and when participants looked at the screen, the experimenter triggered the start of the trial. The image of the condition task appeared at the top of the screen (e.g., eyes for the LOOK condition). The experimenter then pressed the keyboard to trigger the stimulus image, which appeared at the bottom of the screen (e.g., *pig*). After 3 seconds the slide automatically transitioned to a blank screen. On LOOK trials, the stimulus image appeared in silence. On HEARD trials, an auditory token of the word played simultaneously with the appearance of the stimulus image. On PRODUCED trials, the stimulus image appeared in silence and the child produced the word.

After all the training trials were presented, a GIF image of a counting down analog clock appeared on the screen to indicate the start of the break period, during which children were engaged in a block building task. After three minutes, the timer went off which indicated the start of the recall task.

Procedure

Children were individually tested in one session, lasting approximately 10 minutes. Children were randomly assigned to one of six orders that counterbalanced the block order of the three training conditions (LOOK, HEARD, PRODUCED). The experimenter introduced a bear named Mr. Wiggles, who was very tired and going to sleep while the experimenter and participant played a game. Children were told that they were going to learn words, and when Mr. Wiggles woke up they would be asked to tell him the words they had learned. Children were presented with 3 blocks, each consisting of 2 practice trials and 10 training trials on a computer display. At the beginning of each block, the image for the training condition was shown (eyes, ear, mouth), and the experimenter explained what the child was supposed to do. On LOOK trials, participants were asked to look at the screen (no audio information presented). On HEARD trials, participants saw the picture on the screen and simultaneously heard a recording of the word corresponding to the picture. For the PRODUCED condition, participants were asked to say aloud the word corresponding to the picture (no audio was presented). Children started with the 2 practice trials, which were repeated if necessary. Once children had mastered the condition task with the training words, they moved onto the 10 training trials.

After training, children played with WEDGIT™ building blocks for 3 minutes, which was timed by the computer. Once the timer went off, children were told that Mr. Wiggles woke up and were asked to teach Mr. Wiggles the words they had learned. Once their initial recall ended, participants were asked to indicate if they did not remember any other words, they were encouraged to think and try to remember as many as they could. If the child was having difficulties or was too shy, the experimenter could ask questions like “Do you remember any words from the game?” “Do you remember any of the things from the game on the computer

screen?". The experiment ended once participants indicated that they did not remember any more words following the extra prompting just described.

Coding

Each session was video recorded and coded off-line for children's responses during training and recall. Each training trial was coded to ensure that the child gave the correct response for each training condition (e.g., saying the word aloud for PRODUCED condition, remaining silent during HEARD and LOOK conditions). Words from the recall tasks were coded for their training condition. Variation was allowed for recalled items, e.g., the responses *teddy bear* for *bear*, *kitty* for *cat*, *plane* for *airplane*, etc., were coded as correct. We excluded words from the recall analysis if they had training errors (e.g., HEARD item that was produced during training), were practice items, and were extra words (words recalled that were not in training). Words that were recalled more than once were counted as a single recall.

Statistical Analysis

For analysis, we used logistic generalized additive mixed-effects models (GAMMs; Wood, 2017), a type of regression model that can include random as well as (potentially nonlinear) fixed effects. Significance of factors is established through model comparisons, and differences across levels of a (significant) factors can also be computed. One of the main advantages of GAMMs is that they consider continuous factors, such as (apparent) TIME or AGE, not necessarily linearly, using penalized splines to avoid overfitting of the data. They are thus robust against Type II errors and can handle missing data points (e.g., if the data set does not contain points for each and every specific AGE). We implemented GAMMs using *R* with the *mgcv* package (Wood, 2017), and model comparisons (chi-square on maximum-likelihood scores

of two models) as well as difference curves assessing within-factor level differences through TIME (or in the current case, AGE) with *itsadug* (van Rij, Wieling, Baayen & van Rijn, 2017).

We were primarily interested in the effects of WORD-TRAINING condition, AGE (in months), but also included the effect of BLOCK ORDER presentation overall to ensure that there were no effects or interactions of block recency or primacy on word recall (Murdock, 1962; Tan & Ward, 2000). For determining which random effects structure to use for analysis, a (maximal) model was fitted with flat random intercepts by PARTICIPANTS and ITEMS. A second model was fitted to the data using random flat intercepts by PARTICIPANTS and random linear slopes by ITEMS through AGES. Since each participant contributed only one AGE value, no random slope or smooth could be included per PARTICIPANT by AGE, thus flat intercepts were included in all models for participants. Finally, a third model was built with flat intercepts by PARTICIPANTS and nonlinear smooths by ITEMS through AGES. Pairwise model comparisons on all three possible combinations (chi-square of maximum-likelihood scores) determined that the best fit included nonlinear random smooths by AGES and ITEMS, balancing fit and model complexity. Consequently, the statistical model presented below includes random intercepts per PARTICIPANTS and random nonlinear smooths by ITEMS and AGES. Once the random effects structure had been established, we proceeded to determine if BLOCK ORDER was significant or not (comparison of the maximal model with another that did not contain BLOCK ORDER). The best model was refitted using restricted maximum-likelihood method, following Porretta, Kyrölaïnen, van Rij and Järvikivi (2017), for the current presentation.

Results

As mentioned above, in Experiment 1, we were interested in the impacts of WORD-TRAINING condition and AGE (in months) on number of recalled words, in addition to BLOCK

ORDER presentation. Importantly, we modeled AGE as a continuous variable, enabling us to determine if the number of recalled words changed as children aged depending on the TRAINING condition. Number and rate of recalled words by TRAINING condition through apparent time (i.e., AGE in months) are shown in Figure 1. A GAMM comparison (maximum-likelihood scores) between the maximal model (containing all three factors of interest) and a model containing only TRAINING condition and AGE revealed that there was no evidence in favour of the most complex models (Δ in ML score (Δ *df* = 17) = 9.7, $p = 0.31$). This shows that BLOCK ORDER was not significant in Experiment 1, and thus this factor was rejected from the model for further analyses.

The numerical output of the final GAMM is presented in Table 1. One way of visualizing this output is by computing difference curves between two levels of a factor through apparent time. Here, Figure 2 shows all three possible difference curves (i.e., between TRAINING conditions through AGES). The leftmost panel shows that between 36 and 65.9 months old, children recalled significantly more words that were trained in the HEARD condition than in the LOOK condition (green-dotted curve minus blue-dashed curve in Figure 1). In the center panel, model results show that children aged between 36 and 76.8 months old recalled significantly more words that were PRODUCED than LOOKED (red-full curve minus blue-dashed curve in Figure 1). These two results suggest that hearing or pronouncing familiar words, as opposed to looking-only, contributes to higher number of recalls in children early in development, and that this advantage weakens with age. The rightmost panel shows that there were significantly less words trained in the HEARD condition that were recalled than in the PRODUCED condition (green-dotted curve minus red-full curve in Figure 1), which is consistent with the 'classic' production effect, significant for children aged 75.4 months or older. Combining this observation with the tendency observed in the raw data in Figure 1, the evidence indicates a shift of the production effect,

where children on average recalled more HEARD words than PRODUCED word closer to 36 months old, and inversely later in development after 75 months old.

Discussion

Our results showed a reverse production effect for younger children (HEARD > PRODUCED), which shifted to a production effect for the oldest children in our sample (PRODUCED > HEARD). This shift is in line with our prediction that the direction of the production effect is mediated by development related factors. Our results also replicate and extend previous findings from the production effect using a mixed-training design (Icht & Mama, 2015) to a block-training design. Lastly, we observed that early in development, words from the LOOK condition had fewer recalls compared to the HEARD and PRODUCED conditions, though this advantage weakened with age.

Experiment 2

Our initial aim in Experiment 1 was to investigate the production effect across development, starting with children aged 2-years; however, the task was too difficult. When 2-year-olds were tested on the Experiment 1 methodology, there were high attrition rates and many children did not recall any words. In Experiment 2, we adapted our design for 2-year-olds, to verify whether the pattern of the reverse production effect observed with the younger age group in Experiment 1, would be extended to an even younger age-group. To do this, we simplified the task by testing 2-year-old's recall after each of the three training blocks.

Method

Participants

Participants were 30 children aged 2-years (16 males, 13 females, M = 30 months, SD = 3.2 , *range* 24–35). As in Experiment 1, testing occurred in one session. All children were tested

in a museum-based lab and recruited from the museum floor. As in Experiment 1, all children were monolingual speakers of English (*M* = 94%, *SD* = 7.9 *, range* 73–100), with normal hearing, no diagnosed language or developmental delays, as estimated from parental reports. Thirty-seven additional children were tested but not included in the analyses for the following reasons: could not complete the task properly (*n* = 8), no video for off-line coding (*n* = 2), no recalls (*n* = 7), did not complete the experiment (*n* = 14), and parental interference (*n* = 6).

Stimuli

Same stimuli as Experiment 1.

Design

In Experiment 1, after training, children had three minutes to play a puzzle task, followed by the testing phase in which they were asked to recall words from all training conditions. In contrast, in Experiment 2, recall was probed immediately at the end of each of the three training blocks. All other aspects of the design were the same as in Experiment 1.

Procedure

The procedure replicated that of Experiment 1, with the exception that Mr. Wiggles was woken up after each block to test children's recall. No three-minute break was used as the recall task was performed directly after each training block.

Coding

Coding was the same as Experiment 1.

Statistical Analysis

For Experiment 2, we were interested in word-training and block order only, without considering age (in months) in this group. Only one random effects structure was thus possible (i.e., no possibility of inclusion of slopes or smooths through age), thus included flat random

intercepts by PARTICIPANT and ITEM. We started by fitting a "maximal" model with both fixed factors (i.e., WORD-TRAINING condition and BLOCK ORDER). Then, a second model was fitted without the BLOCK ORDER factor and was compared with the maximal model to determine if BLOCK ORDER was significant or not.

Results

Results of Experiment 2 are presented in Figure 3, which displays individual trials (one data point represents one word that was recalled or not) and rates of recalled words for 2-year-olds. Model comparisons between the maximal GAMM (containing both TRAINING condition and BLOCK ORDER) and the second model containing only TRAINING condition did not show any evidence in favour of the most complex model (Δ in ML score (Δ $df = 5$) = 0.93, $p = 0.87$), suggesting that BLOCK ORDER was not significant. Because we were interested in differences across levels of TRAINING condition, this factor was kept in. Results of the parametric section of the final model (see Table 2) revealed that neither HEARD words nor LOOK words yielded significantly more recalls than the PRODUCED words, although the mean recall for HEARD words was greater than PRODUCED words and LOOK words.

To compare the results of both experiments and further explore the development of the production effect between toddlerhood and childhood, we combined all children tested across Experiments 1 and 2 (i.e., 24–83 months old), and computed a difference score for each child, referred to as a Production Effect Score. This difference score was composed of the number of recalled words that were trained in the PRODUCED condition, minus the number of recalled words that were trained in the HEARD condition. A negative score indicates a reverse production effect (i.e., more recalled words that were heard-only than produced), while a positive score indicates a classic production effect (i.e., more recalled words that were produced than heard-only). This

Production Effect Score was submitted to a Pearson's correlation test to assess if child age (in months) had an impact (Figure 4). We found a positive correlation coefficient (r = 0.24) which was significantly different from 0 ($t(148) = 2.98$, $p = 0.003$), indicating that the shift in the direction of the production effect was significant within the group of 24–83 months old participants. Indeed, as shown in Figure 3, values of the Production Effect Score tend to be negative in younger children and positive in older children, with an overall crossover in direction occurring slightly before five years old. This is consistent with our hypothesis that the direction of the production effect shifts as children advance in age.

Discussion

In Experiment 2, children aged 2 years were trained on familiar words under three different conditions: LOOK, HEARD and PRODUCED. Observation of the raw data suggests a reversal of the production effect, i.e., highest recall for HEARD words, although no significant differences emerged following statistical analysis. Of the 30 children tested, 19 recalled more HEARD than PRODUCED words, 3 had equal recalls for HEARD and PRODUCED words, and 8 recalled fewer HEARD than PRODUCED words. The lack of a statistically significant effect could be due to reduced power in Experiment 2 or to overly conservative statistical models. Still, the data are consistent with a shift in the production effect, displayed as a reverse production effect in 2-year-old children. Combining the results from both experiments supports the proposal of a developmental approach to the production effect, in which for younger children production does not result in the same recall benefits for older children.

General Discussion

The current research investigated the production effect across development by testing children whose ages range from 2–6 years on a recall task. In doing so, we aimed to reconcile the

variable results in the handful of existing child-based studies that were obtained using a variety of methods. Our results show that the production advantage is replicable, but is more fragile than initially believed. Findings from Experiments 1 and 2 indicated a relationship between age and the direction of the production effect. The older children are, the more likely they are to show better recall for words they PRODUCED, while younger children show the opposite pattern, with better recall for HEARD words. Words from the LOOK condition were recalled the least, though the difference between the LOOK compared to the HEARD and PRODUCED conditions weakened with age. Overall, the direction of the word recall pattern for younger children was HEARD > PRODUCED > LOOK, while for older children was PRODUCED > HEARD > LOOK. In general, recall rates in our study were lower compared to Icht and Mama (2015). For example, in our study the 5-year-olds recalled an average of 1.7 words from the PRODUCED condition, whereas 5-year-olds recalled about 2.9 words from the PRODUCED condition in Icht and Mama. One possibility is that our computer task was less engaging than their cue-cards task. However, one advantage of our methodology is that the timing and amount of exposure for each auditory and visual stimulus was controlled (rather than variable from one item to another). Below, we first consider the finding for PRODUCED versus HEARD words, then discuss the findings for the LOOK words.

The developmental-based shift in the direction of the production effect for PRODUCED and HEARD words provides support for the cognitive-demand hypothesis, which in combination with the distinctiveness account, gives a more complete picture of what drives the production effect. The current results are consistent with previous work proposing a link between cognitive demands and the influence of production in early development (Vihman et al., 2014; Vihman, 2017). The findings are also consistent with Zamuner et al. (2018), who observed a reverse production effect when older children are trained on auditory novel words paired with visual

nonce animals. The authors had hypothesized that the reverse production effect stemmed from the difficulty of the task. Children did not have existing lexical representations for the novel stimuli; thus, engaging the production system during learning made it more challenging to map the novel word form and referent. As a result, word-learning was detrimentally influenced in the PRODUCED condition, leading to lower recognition compared to the HEARD condition, in which there were more resources available to create new lexical representations. In the current experiments, we found that depending on the age of the child, production can also have an adverse impact on the recall of known words. This supports the idea that higher cognitive demands lead to a reversal of the production effect, and extends the results to factors other than engagement of the production system during novel word learning. Although our stimuli are familiar to all children including the younger ones (produced on average by 84% of 24-month-olds), linguistic knowledge and experience accumulates with age (Werker & Curtin, 2005). Thus, one possible explanation for the reverse production effect in younger children is that the stimuli were less familiar, and their corresponding lexical representations were not well established. Previous research with adults has shown similar learning disruptions with unfamiliar stimuli (Baese-Berk & Samuel, 2016; Cho & Feldman, 2016; Kaushanskaya & Yoo, 2011). If the words were less familiar to the younger children, more attention would have been needed to encode the words, which with the addition of an active task like production, could have led to a disruption in recall. In contrast, for older children, the benefit of production emerged as the words had well-established lexical representations, which subsequently made the task less cognitive demanding, allowing them to benefit from the additional form of encoding. This possibility could be tested by manipulating word frequency: the reverse production effect in younger children should be stronger with lower (than higher) frequency words and the production effect might even be

weaker in older children for very low frequency words.

Relatedly, the age range we studied (our youngest participant was 24-months-old, while the eldest was 83-months-old) was wide enough to span across a child who is just learning how to speak, to an older child who is proficient in comprehending and producing language, allowing us to detect possible effect of language proficiency. Thus, it is possible that the speech production task added a potential layer of task complexity in the current experiments. This could be tested by controlling the words for articulatory complexity: the reverse production effect should be stronger with words that have a higher articulatory load. While our proposal that production can disrupt learning in younger children might seem at odds with predictions from the distinctiveness account because fewer encoding processes in the HEARD conditions should make it harder to recall the HEARD words, language skills go from receptive to expressive knowledge as language develops, and this too may contribute to successful recall. Words first occur in learners' passive lexicon before moving into the active (productive) lexicon, termed the comprehension-expression gap (Schneider, Yurovsky, & Frank, 2015). Various factors influence the transition of a word from the passive to active lexicon (Stokes, de Bree, Kerkhoff, Momenian, & Zamuner, 2019). For instance, words with a higher frequency have a greater probability of being in the active lexicon, as do words with higher phonological neighbourhood density values. Retrieval and assembly of frequent and dense words is faster and more accurate, suggesting that not only will the production effect interact with the words' familiarity, but also phonological density.

Turning to the LOOK words, we now evaluate the possible predictions that were made on how LOOK items might pattern if the younger children recalled more HEARD than PRODUCED words. One proposed possibility was that recall would be the greatest for LOOK words, under the hypothesis that silent observation is the least cognitively demanding task. This was not borne out

by the data, and instead the pattern of recall was HEARD > PRODUCED > LOOK. While the lowest recall for LOOK words is consistent with the distinctiveness account (silent looking involves only one encoding process), this cannot account for the concurrent reversal of performance on recall for HEARD versus PRODUCED words. Still, we also had hypothesized that cognitive resources might be more taxed for younger children having to retrieve the word's phonological representation from memory. While this need to retrieve the word's phonological representation was also true for the PRODUCED condition, it may be that silent retrieval in the LOOK condition also created a less robust memory encoding as compared to retrieving a word and producing it aloud (in which case the auditory feedback as the child produces the word could strengthen the encoding). Lastly, children may have been less engaged/less attentive on the LOOK trials, leading to lower recall; however, in all conditions the experimenter controlled the presentation of the images to when children were looking at the screen.

One important consideration is the equivalence (or lack thereof) of the testing procedures across the production-effect literature: Pritchard et al. (2019) compared the recall of printed words read aloud versus printed words read silently, while the other child-based studies used visual images in LOOK , HEARD or PRODUCED conditions. One could argue that the LOOK condition (silent observation) is equal to Pritchard et al.'s silent reading condition. Looking at Figures 1 and 2, children recalled more words from the PRODUCED condition than LOOK condition (with significant differences up until 6 ½ years), which would then be a replication of Pritchard et al.'s finding of better performance on a yes/no recognition task for words read aloud relative to words read silently, with school-aged children. Perhaps a more analogous reading-based comparison to the current study would be to test children's recognition and recall for printed words read aloud (PRODUCED) versus read aloud by someone else (HEARD) versus read silently

(LOOK). We predict that for children with emerging or less proficient literacy skills, hearing someone read a word aloud would be a less cognitively demanding task, and that recall performance would follow the same pattern as with our youngest participants (HEARD > PRODUCED > LOOK).

To conclude, our findings help reconcile the variable results on the production effect by demonstrating that development plays a role in the effect's direction. When using the same task across a wide-age range, younger children showed a reverse production effect, and as children grew older, they became more likely to display a production advantage. Although we characterized the shift in the production effect as age-dependent, we did not include any additional measures of processing, such as children's language skills, executive functioning, and working memory capacity. As these skills are inter-related (e.g., Gathercole, Willis, Emslie, & Baddeley, 1992; Swanson, 1996**)**, future work with additional measures is needed to better understand what drives shifts in the direction of the production effect. Nevertheless, taken together with previous results in the literature, the current studies remind us that multiple factors can, and almost always, interact across different levels of development; moreover, they provide a more complete understanding of the complex effect of the role of production on language learning.

References

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

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

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(44), 13531–13536. https://doi.org/10.1073/pnas.1508631112

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*, 1-11. https://doi.org/10.1080/20445911.2016.1182173

Dahlen, K., & Caldwell-Harris, C. (2013). Rehearsal and aptitude in foreign vocabulary learning. *The Modern Languages Journal, 97*, 902-916. https://doi.org/10.1111/j.1540-4781.2013.12045.x

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

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 534-537.

Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur communicative development inventories: User's guide and technical manual*. Singular.

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

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677. https://doi.org/10.1017/S0305000916000209

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*(5), 887–898. https://doi.org/10.1037/0012-1649.28.5.887

Icht, M., & Mama, Y. (2015). The production effect in memory: A prominent mnemonic in children. *Journal of Child Language*, *42*, 1102-1124. DOI: 10.1017/S0305000914000713

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

Kan, P. F., Sadagopan, N., Janich, L., & Andrade, M. (2014). Effects of speech practice on fast mapping in monolingual and bilingual speakers. *Journal of Speech, Language, and Hearing Research*, *57*(3), 929-941. DOI: 10.1044/2013_JSLHR-L-13-0045

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

Krishnan, S., Watkins, K. E., & Bishop, D. V. (2017). The effect recall, reproduction, and restudy on word learning: A pre-registered study. *BMC Psychology*, *5*(1), 1-14. DOI: 10.1186/s40359-017-0198-8

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology*, *55*(4), 306-353. https://doi.org/10.1016/j.cogpsych.2007.01.001

MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review*, *18*, 1197-1202. https://doi.org/10.3758/s13423-011-0168-8

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,* 671-685. https://doi.org/10.1037/a0018785

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

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64*(5), 482–488. https://doi.org/10.1037/h0045106

Ngon, C., & Peperkamp, S. (2016). What infants know about the unsaid: Phonological categorization in the absence of auditory input. *Cognition*, *152*, 53-60. https://doi.org/10.1016/j.cognition.2016.03.014

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory*, *22*, 509-524. https://doi.org/10.1080/09658211.2013.800554

Porretta, V., Kyrölaïnen, van Rij, J., & Järvikivi, J. (2018). Visual World Paradigm data: From preprocessing to nonlinear time-course analysis. In I. Czarnowski, R. J. Howlett, & L. C.

Jain (Eds.), *Intelligent Decision Technologies 2017: Proceedings of the 9th KES International Conference on Intelligent Decision Technologies – Part II* (pp. 268-277). Springer International Publishing.

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

Richtsmeier, P., & Moore, M. (in press). Order effects in the perception and production of new words. *Journal of Speech, Language, and Hearing Research.*

Schneider, R., Yurovsky, D., & Frank, M. (2015). Large-scale investigations of variability in children's first words. In D. C. Noelle, Dale, R., Warlaumont, A. S., Yoshimi, J., Matlock, T., Jennings, C. D., & Maglio, P. P. (Eds.), *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2110-2115): Cognitive Science Society. Retrieved from https://cogsci.mindmodeling.org/2015/papers/0364/

Stokes, S. F., de Bree, E., Kerkhoff, A., Momenian, M., & Zamuner, T. (2019). Phonology, semantics, and the comprehension–expression gap in emerging lexicons. *Journal of Speech, Language, and Hearing Research*, *62*(12), 4509-4522. DOI: 10.1044/2019_JSLHR-19-00177

Swanson, H. L. (1996). Individual and age-related differences in children's working memory. *Memory & Cognition*, *24*(1), 70-82. DOI: 10.3758/bf03197273

Tan, L., & Ward, G. (2000). A recency-based account of the primacy effect in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*(6), 1589–1625. https://doi.org/10.1037/0278-7393.26.6.1589

van Rij, J., Wieling, M., Baayen, R. H., & van Rijn, H. (2017). *itsadug: Interpreting time series and autocorrelated data using GAMMs* [R package]. Retrieved from https://rdrr.io/cran/itsadug/man/itsadug.html

Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology*, *108*(1), 1–27. https://doi.org/10.1111/bjop.12207

Vihman, M. M, DePaolis, R. A., & Keren-Portnoy, T. (2014). The Role of Production in Infant Word Learning. *Language Learning*, *64*, 121–140. https://doi.org/10.1111/lang.12058

Werker, J.F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development, 1*, 197-234. https://doi.org/10.1207/s15473341lld0102_4

Wood, S. (2017). *Generalized Additive Models: An Introduction with R* (2nd ed.). Boca Raton: Chapman and Hall/RC.

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *245*, 603–612. https://doi.org/10.1177/0956797612458802

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

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). https://doi.org/10.1111/desc.12636

Zamuner, T. S., & Thiessen, A., (2018). A phonological, lexical, and phonetic analysis of the new words that young children imitate. *Canadian Journal of Linguistics*, *63*, 609-632. DOI: https://doi.org/10.1017/cnj.2018.10

Table 1

Numerical output of the GAMM for recall data in children aged three to six years.

Formula: recall ~ TRAINING + s(AGE) + s(AGE, by=TRAINING) + s(PARTICIPANT, bs="re") + s(AGE, ITEM, bs="fs", m=1)				
Parametric coefficients	Estimate	Std. Error	Z value	p-value
Intercept(TRAINING:PRODUCED)	-1.905	0.123	-15.502	< 0.001
TRAINING:HEARD	-0.058	0.124	-0.472	0.637
TRAINING:LOOK	-0.49	0.138	-3.543	< 0.001
Smooth terms	edf	Ref.df	χ^2	p-value
s(AGE)	1	1.001	1.062	0.303
s(AGE, by= TRAINING:PRODUCED)	1.001	1.001	4.134	0.042
s(AGE, by= TRAINING:HEARD)	0.0003	0.0005	0	0.994
s(AGE, by= TRAINING:LOOK)	1	1.001	6.835	0.009
s(PARTICIPANT) (RANDOM INTERCEPT)	1.911	118	1.937	0.441
s(AGE:ITEM) (RANDOM SMOOTH)	23.06	269	72.608	< 0.001

Table 2

Numerical output of the GAMM for recall data in children aged two years.

Formula: recall ~ TRAINING + s(PARTICIPANT, bs="re") + s(ITEM, bs="re")				
Parametric coefficients	Estimate	Std. Error	Z value	p-value
Intercept(TRAINING:PRODUCED)	-1.988	0.241	-8.247	< 0.001
TRAINING:HEARD	0.394	0.235	1.676	0.094
TRAINING:LOOK	-0.083	0.253	-0.327	0.744
Smooth terms	edf	Ref.df	χ^2	p-value
s(PARTICIPANT) (RANDOM INTERCEPT)	14.66	29	29.92	0.002
s(ITEM) (RANDOM INTERCEPT)	17.56	29	45.66	< 0.001

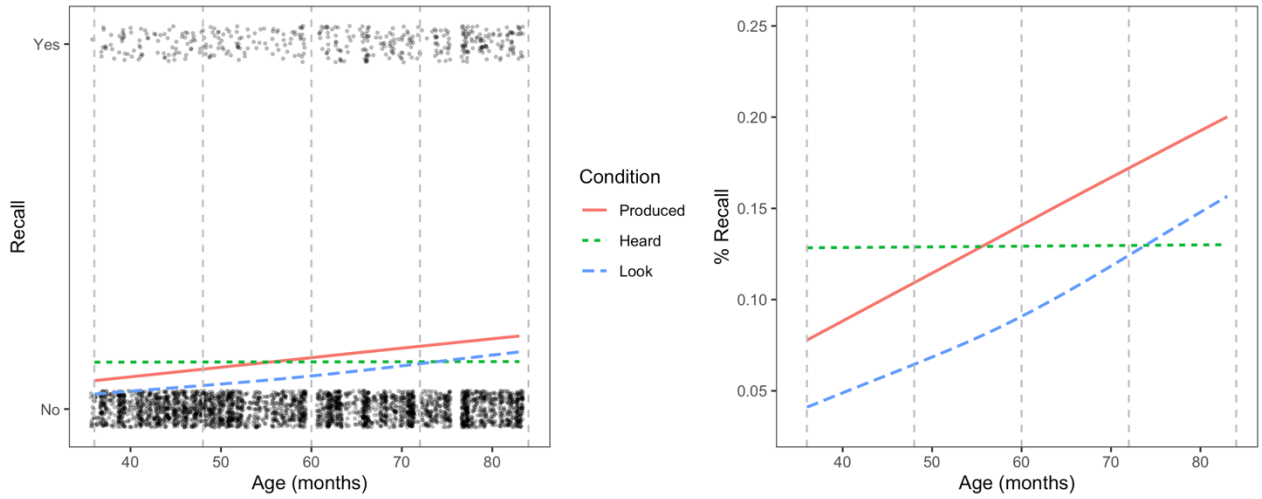

Figure 1. Effect of TRAINING condition and AGE (months) on number of recalls for three- to six-year-old children in Experiment 1. Grey clusters of dots represent individual trials, in which a word was either recalled (Yes) or not (No). Coloured lines represent the mean proportion of recall for each TRAINING condition. Note that the right panel shows the same data as the left panel, showing only the average rates of recall through apparent time (AGE).

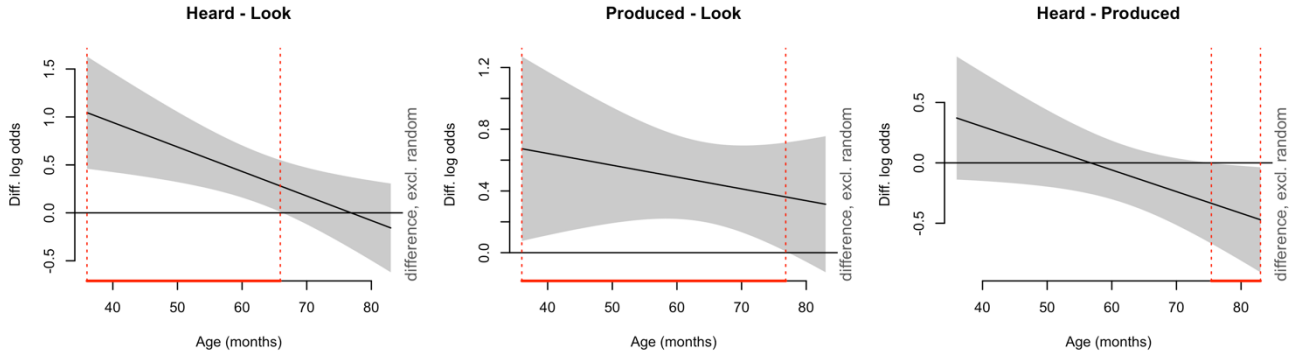

Figure 2. Difference curves (GAMM output) for probability of recall through AGE by WORD-TRAINING condition. Significant intervals are displayed in red between dotted lines (i.e., where the 95% confidence interval does not overlap with 0).

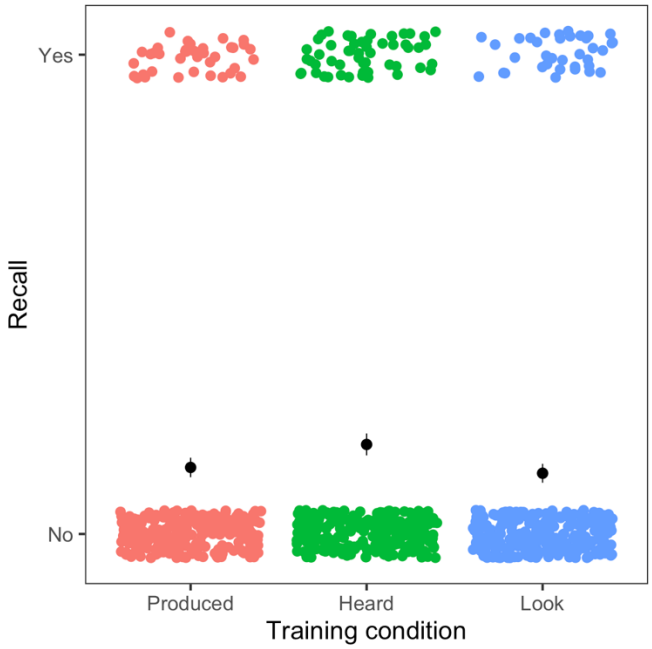

Figure 3. Effect of TRAINING condition on number of recalls for two-year-old children in Experiment 2. Coloured clusters of dots represent individual trials, in which a word was either recalled (Yes) or not (No). Black dots represent the mean proportion of recall for each TRAINING condition.

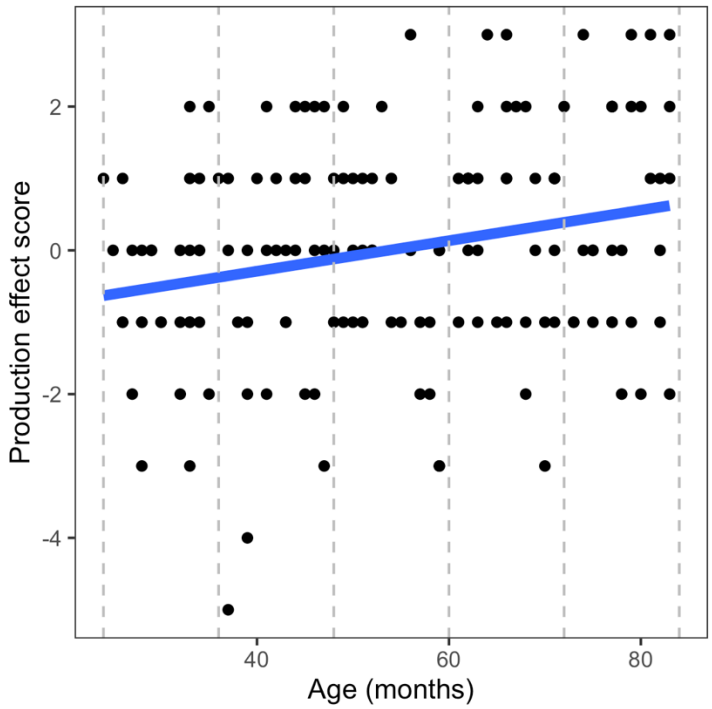

Figure 4. Production Effect Scores by age (in months), i.e., number of recalled words that were produced minus that were heard-only, for all children in Experiments 1 and 2, and regression line (linear).