Reconciling the contradictory effects of production on word learning: Production may help at first, but it hurts later

Efthymia C. Kapnoula

Basque Center on Cognition, Brain and Language; Ikerbasque

and

Arthur G. Samuel

Basque Center on Cognition, Brain and Language; Stony Brook University; Ikerbasque

Running Head: PRODUCTION HELPS AT FIRST, BUT HURTS LATER

Corresponding Author:
Efthymia C. Kapnoula
Paseo Mikeletegi 69
Basque Center on Cognition, Brain and Language (BCBL)
20009 San Sebastián - Donostia
Spain
kapnoula@gmail.com

Abstract

Does saying a novel word help to recognize it later? Previous research on the effect of production on this aspect of word learning is inconclusive, as both facilitatory and detrimental effects of production are reported. In a set of three experiments, we sought to reconcile the seemingly contrasting findings by disentangling the production from other effects. In Experiment 1, participants learned eight new words and their visual referents. On each trial, participants heard a novel word twice: either (1) by hearing the same speaker produce it twice (*Perception-Only condition*) or (2) by first hearing the speaker once and then producing it themselves (*Production condition*). At test, participants saw two pictures while hearing a novel word and were asked to choose its correct referent. Experiment 2 was identical to Experiment 1, except that in the Perception-Only condition each word was spoken by two different speakers (equalizing talker variability between conditions). Experiment 3 was identical to Experiment 2, but at test words were spoken by a novel speaker to assess generalizability of the effect. Accuracy, RT, and eye-movements to the target image were collected. Production had a facilitatory effect during early stages of learning (after short training), but its effect became detrimental after additional training. The results help to reconcile conflicting findings regarding the role of production on word learning. This work is relevant to a wide range of research on human learning in showing that the same factor may play a different role at different stages of learning.

Keywords: word learning, spoken word recognition, production, mental lexicon, visual world paradigm

Introduction

To learn a novel word, we need to integrate it into our mental lexicon. The trajectory of lexical integration and the factors upon which it depends are hotly debated topics. Here we examine the role of *production* (i.e., the action of producing a word out loud compared to just hearing it) on word learning, using spoken word recognition as a proxy of lexical integration.

There is a widely held assumption that producing a novel word helps to build its lexical representation, which can then support lexical functions, including its recognition. In accordance with this assumption, when people learn a new word, they are often asked to repeat it. This is a common practice, for example, in second language learning contexts, where instructors often ask their students to repeat new words immediately after encountering them for the first time (Duff, 2000; Kadota, 2019). In line with this idea, current approaches to second language learning, such as Communicative Language Teaching, Task-Based Language Teaching, and Communicative Competence emphasize learners' immediate communication needs and encourage production from the earliest moments of instruction (for a review see Lightbown & Spada, 2013).

The idea that production helps word learning is in line with a substantial body of research showing that production enhances memory. First reported by Hopkins and Edwards (1972), the key finding is that material read aloud is better remembered. This facilitatory *production effect* has been documented many times. For example, Gathercole and Conway (1988) conducted a series of experiments in which adults were presented with a set of words (one by one) and, depending on the experiment, were asked to read a word silently, read it out loud, mouth it, read it and hear it, only hear it, or write it (seeing it or not). At test, participants had to indicate whether a given word was new or old. Across the experiments, performance at test was better for words that were read out loud compared to all other conditions (see also Dodson & Schacter, 2001; Gathercole & Conway, 1988; P. MacDonald & MacLeod, 1998; MacLeod & Bodner, 2017).

Given the robustness of this phenomenon, it seems intuitively reasonable to assume that similar facilitation should apply to novel words; however, research on this topic has been inconclusive. There is some evidence that production facilitates word learning, but there is also evidence that production can actually impair word learning. The difficulty in assessing the role of production in word learning stems mainly from the fact that very few studies have tried to isolate the effect of production independently of other effects, such as the testing effect (Karpicke & Roediger, 2008), the finding that recall from memory benefits the retention of novel information.

To date, studies that have looked more closely at the effect of production have reported contradictory results. For example, Zamuner et al. (2016) report evidence from eye-movements that production has a facilitatory effect on novel word learning. However, Zamuner et al. (2018, using the same task, but with children rather than adults) reported the reverse pattern, i.e., a detrimental effect of production. Similarly, Leach and Samuel (2007) also showed that production during training leads to weaker effects of lexical engagement, meaning that produced words were not as well integrated into the mental lexicon, compared to words that were only heard during training.

The present study investigates the source(s) of these discrepancies to shed light on the role of production on the learning of novel spoken words, as reflected by their recognition. Given the existence of both positive and negative consequences of production, we begin by briefly presenting a number of mechanisms via which production may facilitate or impair word learning. Our goal was to identify possible confounds and isolate the direct effect(s) of production per se.

Reasons why production may help word learning

As mentioned above, even though there is evidence consistent with a facilitatory role of production in word learning, it is difficult to assess whether production per se is the critical factor behind these effects and, more importantly, to pinpoint the exact mechanism. Work on the production effect on familiar words has generated support for a *distinctiveness* account, according to which, the act of producing a word enhances its distinctiveness and thus strengthens the corresponding mnemonic trace (Gathercole & Conway, 1988; MacLeod & Bodner, 2017; Ozubko et al., 2014; Ozubko & Macleod, 2010). However, learning a novel word involves more than just generating a mnemonic trace; it requires creating an entirely novel lexical representation and integrating it into the mental lexicon. Thus, it is unknown whether and how production may help different aspects of word learning. Below we describe a few different mechanisms via which this facilitation may occur.

First, production may boost word learning via the creation of *articulatory representations*. According to the model of speech production proposed by Hickok and colleagues (Hickok, 2012, 2014; Hickok et al., 2011), phonological representations act as hubs that are used to map sound to meaning and to speech articulation. Even though the two mappings are computationally distinguishable (Nora et al., 2015), they can both be viewed as components of lexical representations. Thus, learning and practicing the articulatory sequence that corresponds to a novel word may serve as an additional dimension of the newly formed lexical representation, which can then be used to bootstrap further integration of the word into the mental lexicon. An articulatory locus for the facilitatory effect of production would also be consistent with work on second-language acquisition showing a learning advantage for overt but not covert repetition (Mattys & Baddeley, 2019). This mechanism could be considered a facilitatory effect of production per se.

Second, it has been proposed that word learning is tightly linked to and supported by *phonological short-term memory* (PSTM; for review, see Baddeley, Papagno, & Vallar, 1988; Gathercole, 2006). The general idea is that in order to produce a word, the corresponding phonological sequence is briefly maintained in PSTM, which gradually leads to longer-term learning. However, the details of this mechanism vary across different models (for a review see Thorn & Page, 2008). For example, according to the primacy model, proposed by Page and Norris (Page & Norris, 1998, 2008, 2009), phonological word-form learning is a more naturalistic version of the Hebb repetition effect (Hebb, 1961; which refers to the finding that immediate serial recall of a list of familiar items, such as digits, gradually improves over multiple repetitions). On the other hand, Gupta's model (Gupta, 2003, 2008; Gupta & Tisdale, 2009) involves a short-term sequencing mechanism that builds associations between sublexical sequences and patterns of lexical-level activation. In that respect, Gupta's model makes a distinction between sublexical sequences and lexical word-forms. In this case the facilitatory effect would again be inherently linked to production.

Third, a broader mechanism that may be responsible for the facilitatory effects of production is *attention*. The idea that language production requires higher levels of attention is not only intuitive, but also supported by data (e.g., see Boiteau, Malone, Peters, and Almor, 2014). Despite this, few experiments have controlled for this factor in assessing the role of production in word learning. For example, in the Gathercole and Conway (1988) experiments mentioned above, words were either preceded by the critical instruction (e.g., "say aloud"), or they were only presented in one experimental condition for a given participant. Thus, in both cases participants knew if they were required to produce the word ahead of time. Knowing that they will need to say the target word out loud may have led to increased attention – e.g., towards the phonological structure of the word, its acoustic implementation, its possible semantic associations, or any combination of these – which could explain the robust facilitatory effect of production during testing. Thus, any facilitation caused by attention would be an indirect consequence of production.

Fourth, often (but not always[1]) production involves recall from memory. This is the case, for example, when a novel word-form is linked to a visual referent, which is subsequently (at training and/or at testing) used to prompt the production of its newly learned label. Retrieval practice is known to lead to better retention of information (i.e., the *testing effect*; Roediger & Karpicke, 2006). In support of this hypothesis, Karpicke and Roediger (2008) trained English-speaking adults in Swahili-English word pairs using training regimes that differed in whether they involved repeated testing (involving recall) versus repeated studying (not involving recall). In contrast to repeated studying, which had no effect on delayed recall, repeated testing had a large facilitative effect. Similarly to the previous case, this kind of facilitation would be viewed as a by-product of production.

Lastly, there is an additional way in which production can differ from passive exposure. In training paradigms that involve auditory presentation of the target words, when participants are asked to produce a word themselves, they also hear it in a new voice (their own). In these cases, production is confounded with increased *speaker variability*. When learning new words, listeners encode voice-related information (Creel & Tumlin, 2011; Houston & Jusczyk, 2000; Kapnoula & Samuel, 2019), which is why variability in this dimension may affect word learning.

[1] When production immediately follows presentation of the target word, no retrieval from long-term memory is necessary.

Indeed, there is evidence in favor of a beneficial role of talker variability in word learning. For example, Rost and McMurray (2009) showed that when infants learn similar words (like *buk* and *puk*) spoken by multiple speakers, they later discriminate between them better than when they have only heard the words spoken by one person (see also Höhle et al., 2020). Similarly, Richtsmeier et al (2009) exposed 4-year-olds to novel words that were spoken either by one or 10 different talkers. At test, children were faster and made fewer errors in producing the words that had been spoken by many talkers. These findings suggest that hearing a novel word spoken by different talkers can lead to a more robust and abstract lexical representation. A proposed mechanism behind this effect is that increased variability in irrelevant dimensions helps the listener to identify the relevant dimensions (Rost & McMurray, 2010; Singh, 2008). In this case, facilitation would be caused by the additional variability that comes with production, rather than production itself.

Based on the points outlined above, it is clear that including a production requirement in a training regime may lead to better word learning, but, depending on the details of the procedure, this may be due to a number of different mechanisms. Thus, in order to fully understand the role of production in word learning, we need to use experimental designs that take these points into consideration.

Reasons why production may hinder word learning

Although counterintuitive, it is also theoretically possible that production could impede word learning. Again, there are a number of different ways in which this could happen. First, production may interfere with the *encoding* of a novel word-form at the earliest moments of learning. By encoding we refer to learning the sound pattern of a word-form, which can be viewed as the bare minimum amount of information that is necessary to recognize a word. This maps onto what Leach and Samuel (2007) refer to as *lexical configuration* and it is thought to correspond to early stages of word learning (i.e., when the word is first added into the mental lexicon, after a handful of exposures to its spoken form). That is, immediately after hearing the new word, listeners may benefit from having a moment in which no further input (or output) is processed. During this period, the system can make the necessary adjustments (e.g., adjust the connection weights between speech sounds and lexical levels) that correspond to the successful encoding of the novel word-form. Specifically, production may hinder this process by blocking access to the echoic trace of the stimulus. This may in turn impede encoding directly, by taking away the input of encoding, and/or indirectly, by taking away the input of sub-vocal rehearsal, which would be expected to boost encoding. In these ways, immediate production may interfere with the encoding process, yielding non-optimal learning outcomes.

Second, and relatedly, if learners are simultaneously dealing with the need to learn a word perceptually and to learn how to produce it, any mismatch between the memory representations and/or cognitive processes needed for these two tasks can lead to *interference* between them. Given that perception is based on auditory codes and production is based on motor codes, there is inherently some mismatch. The fact that both codes need to refer to the same object at some level means that they may be particularly vulnerable to a form of lateral inhibition. The idea that production may interfere with perception during early stages of learning is not new. Krashen’s Input Hypothesis (1985), for example, makes this point in the context of second language learning. According to this hypothesis, when learning a new language one should not rush into producing new words during the very first stages of language learning. Instead, production should follow after a “silent period” has passed. Even though this is based on developmental observations of L1 acquisition, the idea is applicable to adult language learning.

Third, production may interfere with word learning at a later stage, when the word-form is being linked to its *semantic referent*. Here, the underlying assumption is that a new word-form is first encoded (in some form of proto-lexical representation) and then mapped to a referent (Fernandes et al., 2009; Rodriguez-Fornells et al., 2009, but see François et al., 2017, for evidence that they can also happen in parallel). Indeed, a number of studies have shown that mapping to meaning is facilitated when it follows speech segmentation (Graf Estes et al., 2007; Hay et al., 2011; Mirman, Magnuson, et al., 2008). The rationale and corresponding processes of how production interferes with word learning would be very similar to the kind of interference described above (during encoding), but in this case production would overlap with the (subsequent) mapping of the novel word to its meaning.

Fourth, it might also be the case that production exposes learners to *poor input*, which leads to poorer learning. That is, the input that is presented to the participants during training is usually comprised of clean, carefully manipulated, high-quality stimuli. In contrast, the nature of a learner’s own production is out of the experimenter’s control and can thus vary substantially in terms of quality. Noisy output can act as input, for example when the participant is asked to read a word, or hear and repeat it. As a result of being exposed to noisy input, learning may be negatively affected.

Indeed, the idea that the quality of the input can affect processing of spoken language is intuitive and supported by empirical findings. Relevant work has mostly looked at *clear speech*, which is slower and hyperarticulated compared to plain speech (Bradlow et al., 1996; Bradlow & Hayes, 2003). Speakers typically adopt this speaking style when the listener is thought to face communication-related difficulties, e.g., nonnative language or hearing impairment (Smiljanić & Bradlow, 2009), and indeed clear speech input appears to facilitate comprehension (e.g., see

Payton et al., 1994). More pertinent to our study, Riley and McGregor (2012) examined the effects of speaking style on children's word learning. They found that new words heard in clear speech were later produced more accurately; however no effect of speaking style was found for perception (see also Baese-Berk & Samuel, 2016, for a discussion of the role of input quality in perceptual learning).

Previous research on the role of production in word learning is inconclusive

It should not be surprising that studies on the role of production on word learning have yielded contradictory findings; given the variety of mechanisms in which production may affect word learning, seemingly small differences between experimental designs and procedures may lead to large differences in (or even reversal of) the obtained pattern of results.

Leach and Samuel (2007) evaluated how a number of different factors affected novel word learning. In this work, the authors focused on two lexical properties: *configuration* and *engagement*. In the current context, as mentioned above, lexical configuration corresponds to building a phonological representation. In contrast, lexical engagement refers to the ways in which a word interacts with other representations (e.g., inhibiting other words, or boosting the activation of speech sound representations). The latter property is taken as a stronger marker of word learning, as it reflects deeper integration of a novel item into the mental lexicon.

In the Leach and Samuel (2007) study, participants were either trained with a phoneme monitoring task (Exp.1) or a word-picture-association task (Exp.2), coupled with a production requirement (Exp.4) or not (Exp.5). After training, lexical configuration and lexical engagement were assessed separately. Lexical configuration was assessed in a three alternative recognition judgment (in which participants would hear a newly learned word along with two similar lures and had to choose which of the three items they had just learned) and a word-in-noise task (in which participants had to recognize the critical items buried in progressively lower levels of white noise). To assess lexical engagement the authors measured the ability of the new items to drive phonemic restoration (Samuel, 1996; Warren, 1970; the finding that when part of a word is missing or replaced by a different sound, listeners still report hearing it) and perceptual learning (Norris, McQueen, & Cutler, 2003; the finding that repeated exposure to an ambiguous sound embedded in real words changes the way listeners identify this sound in a later task). Interestingly, the results revealed a dissociation: production boosted lexical configuration, but hindered lexical engagement.

Hopman and Macdonald (2018) also looked at the role of production, but unlike the Leach and Samuel study, their production task required participants to recall the critical information (i.e., the newly learned words). Word learning was assessed via a vocabulary test, which required the comprehension of individual words within a phrase context. Their results revealed a facilitatory effect of production on word learning. However, it remains unclear whether this effect was due to production per se, or can, for example, be attributed to repeated retrieval (i.e., the testing effect). Additionally, it could be argued that their measure of word learning assessed speed of phrase comprehension, rather than word learning per se.

Finally, Zamuner et al. (2016) looked at the role of production in word learning using eye-tracking. During training each new word was presented along with its visual referent. For half of the items, participants heard the new word twice (Heard-Only condition) and for the other half they heard it once and were required to repeat it themselves (Produced condition), thus equalizing the number of times each item was presented in each condition. At test, participants heard each word and had to select its correct referent, given two options. Eye-movements during testing were analyzed using growth curve analysis (GCA; Mirman, Dixon, & Magnuson, 2008), which revealed a significant difference in the shape of the looking curves (quadratic term) between conditions. The authors interpreted this difference as evidence for a facilitatory effect of production on word learning.

A possible concern is that production was confounded with speaker variability; Heard-Only items were heard twice by the same speaker (and the same recorded token), whereas Produced items were heard by two speakers (the voice played to them and the participant themselves). Thus, it is unclear whether the significant difference between experimental conditions was driven by production or by input variability. In addition, training was limited to two trials per item (which is much lower than the number of training repetitions typically used in word learning studies, e.g., Gaskell & Dumay, 2003; Kapnoula et al., 2015; Leach & Samuel, 2007). This leaves open the possibility that this effect only appears at very early stages of word learning, which may not reflect integration of the new items into the mental lexicon.

Present study

The main goal of the present study is to examine the effect of production on word learning independently of other commonly confounding factors (talker variability, the testing effect, attention, etc.), in order to offer an account that reconciles previous results. In addition, to achieve a more comprehensive understanding of the effect, we looked at whether/how this effect is modulated by the amount of training.

We adopted Zamuner et al.'s (2016) approach, which combines a number of strengths, such as: 1) the number of presentations is equal across conditions, 2) retrieval is not required for production (i.e., production is not confounded with retrieval practice), 3) participants are not instructed about the mode of response ahead of time (i.e., minimizing differences in attention), and 4) eye-movement data can be used as a proxy of lexical activation, allowing us to track lexical activation in real time (Allopenna et al., 1998; Salverda & Tanenhaus, 2017). In addition, by adopting this design we can more directly compare our results to those of the original study.

We conducted three experiments with the same general structure/design, each one focusing on a different question. Experiment 1 aimed at replicating Zamuner et al. and examining whether/how the results change as a function of the amount of training. To test this, we added further training and testing after what corresponded to the end of the Zamuner et al. experiment. Experiment 2 examined the effect of production on word learning while controlling for speaker variability. To achieve this, we introduced speaker variability[2] in the Perception-Only condition by playing each word in two different voices (thus matching the variability present in the Produced condition). In Experiment 3, we used a novel voice at test to examine whether word learning with/without production generalizes differently to novel speakers (e.g., whether production leads to better generalization to novel speakers).

In all three experiments, we used the visual word paradigm (VWP) to track activation of the target word in real time. In this paradigm, the underlying hypothesis is that the probability of looking at an object increases as a function of the activation of the corresponding lexical item. Based on this linking hypothesis, fixation proportions over time can be used as a direct index of lexical activation (Allopenna et al., 1998; Salverda & Tanenhaus, 2017; Tanenhaus et al., 1995; see also Magnuson, 2019 for a review of alternative hypotheses). Across experiments, the effect of Production is defined as the difference between the Production and Perception-Only conditions, since our question was how producing a new word affects learning compared to just hearing it.

[2] Note that in other studies the typical number of talkers in high-variability conditions is much higher than two; however, in contrast to those studies, our goal was to control for talker variability, rather than test its effect.

Experiment 1

Experiment 1 is intended to replicate the facilitatory effect of production on word learning reported by Zamuner et al. (2016). Note that Zamuner et al. (2016) used two training trials per item, which is much lower than the typical amount of training used in word learning studies (11-24 trials; e.g., see Gaskell & Dumay, 2003; Kapnoula et al., 2015; Kapnoula & McMurray, 2016; Leach & Samuel, 2007). Thus, it is possible that the facilitatory effect of production applies specifically to early stages of lexical acquisition (e.g., lexical configuration, if we adopt the terminology proposed by Leach and Samuel). To test this, we asked whether this effect is modulated by the amount of training.

Method

Participants

Forty (31 females; mean age = 25.8 years) native speakers of Spanish participated in Experiment 1. Power analyses were conducted on data from three previous eye-tracking experiments (reported in Kapnoula et al., 2015; Kapnoula & McMurray, 2016) that used a different within-subject manipulation. Given the absence (to our knowledge) of a well-tested method of sample size estimation for curve-fitting analyses (which was our primary analytical approach), we conducted analyses for repeated-measures, within-subjects ANOVA (which was our secondary analytical approach). These analyses indicated that a power of .95 requires a sample size of 31 to 41. All analyses were conducted in G*Power (Faul et al., 2009, 2007).

Most participants were also fluent in Basque, which was foreseen and taken into account in selecting the stimuli (see ***Materials*** below). All participants had normal/corrected-to-normal vision and no known hearing or neurological impairments. Participants underwent informed consent and were remunerated for their participation. All experimental procedures were approved by the BCBL ethics committee.

Design

Experiment 1 had two Phases, each consisting of one training and one testing block (see Table 1). Phase 1 of the training had the same number of training trials used by Zamuner et al., 2016, while Phase 2 had five times that number of trials. Thus, there were 12 repetitions across Phase 1 and Phase 2 training, which matches the typical amount of training used in previous word learning studies (11-24 trials). In such previous studies, learners typically show asymptotic performance after approximately 8-10 trials (e.g., Leach & Samuel, 2007; Samuel & Larraza, 2015). The testing blocks were identical across the two Phases and the test trials in each Phase matched the number of test trials (32) in the Zamuner et al. study.

Table 1. *Number of training and testing trials per phase*

	Phase 1		**Phase 2**	
Training	8 words × 2 repetitions =	16 trials	8 words × 10 repetitions =	80 trials
Testing	8 words × 4 repetitions =	32 trials	8 words × 4 repetitions =	32 trials

Training. Participants learned eight novel words. Each word-form was used as the label for an unfamiliar object. The correspondence between novel words and objects was randomized across participants using a Latin Square.

Crucially, for each participant, half of the words were assigned to the Perception-Only condition, and half to the Production condition (the assignment of each word to one condition or the other was randomized across participants). In Perception-Only training trials, the picture of one unfamiliar object was presented and the corresponding novel word was heard twice. In

Production training trials, the only difference was that the corresponding novel word was heard once and then repeated by the participant.

In Phase 1, we used the same number of training trials used by Zamuner et al. (2016), which was two trials per word (8 words × 2 repetitions = 16 training trials). In contrast, the training block of Phase 2 consisted of 10 trials per word (8 words × 10 repetitions = 80 training trials). Thus, cumulatively, there were 96 training trials across Phases.

Testing. Participants were tested twice, once at the end of each Phase. All testing trials were identical across conditions and phases. In each testing trial, participants saw two of the objects while hearing one of the novel words and had to select the picture that was the correct referent of that word. Each target was repeated four times in each testing block, resulting in 64 testing trials (8 words × 4 repetitions × 2 testing blocks) across Phases.

Materials

All novel words were CVCV items: /bopa/, /ʧofa/, /deɾa/, /guθa/, /kiða/, /reka/, /tuma/, and /jata/. The items were checked by a Spanish-Basque bilingual research assistant to make sure that they were nonwords in both Spanish and Basque, but morphologically consistent with Spanish. Spoken stimuli were recorded by a native female speaker of Spanish in a sound-attenuated room, sampling at 44,100Hz. We collected multiple recordings and chose one recording per item based on sound quality. Chosen recordings were cut, cleaned (background noise and occasional click/pop sounds removed), and intensity-scaled. Finally, 50 ms of silence was added before and after each word. The average duration of the final stimuli (including the 100 ms of silence) was 714 ms.

Visual stimuli consisted of color pictures of eight unfamiliar objects. All images were taken from Horst & Hout (2016). Images measured 300 × 300 pixels during presentation.

Procedure

Participants were seated in front of a computer screen and were instructed that their task would be to learn a set of new words and their meanings. They were then fitted with an SR Research EyeLink 2K eye-tracker, a system with remote desktop mounting.

After calibration, participants were given instructions for the task and did a short training-and-testing practice. The practice stimuli consisted of six images of fruits (piña [pineapple], melon [melon], mango [mango], fresa [strawberry], uvas [grapes], and pera [pear]) and their corresponding names, recorded by the same speaker as the experimental stimuli.

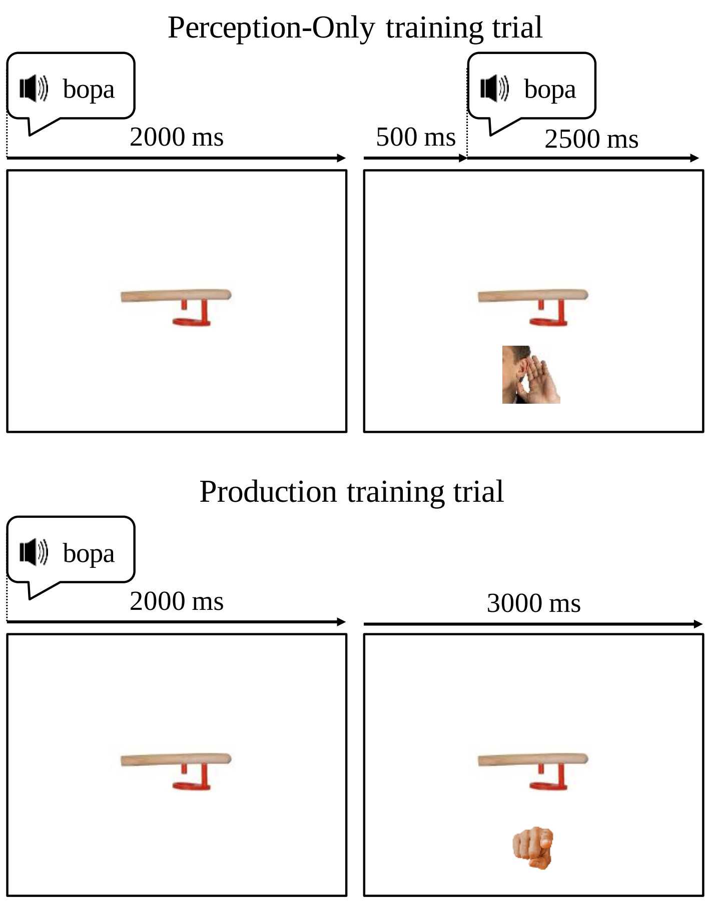

Figure 1. Visual description of training trial per condition

At the beginning of each training trial, one picture was presented at the center of the screen of a 19” monitor operating at a resolution of 1204 × 768 pixels. Simultaneously, the auditory label of the picture was played through high quality headphones. A prompt image appeared below the image of the unfamiliar object 2,000 ms after the onset of the word. In Perception-Only trials, the prompt image showed a hand next to an ear making the gesture for “listen”. When seeing this prompt, participants had to remain silent and listen to the word again. The novel word was repeated by the computer 500 ms after the presentation of the prompt. In production trials, the prompt image showed a finger pointing at the participant. When seeing this prompt, participants had to repeat the word out loud into a microphone attached to their headphones. The prompt images were presented and explained to the participants during the initial instructions. For both conditions, the image of the unfamiliar object remained on the screen for 3,000 ms after the presentation of the prompt (see Figure 1). Perception-Only and Production trials were randomly intermixed. The first training Phase lasted approximately 2-3 minutes and the second one approximately 10 minutes.

At the beginning of each testing trial, pictures of two objects were presented in the two horizontal ends of the screen, spaced 424 pixels apart. One picture was the target item for that trial (i.e., the picture assigned to that word during training). The other picture was one of the other seven images. Items were paired so that each target was always presented with the same competitor. This was done to minimize further learning opportunities during testing[3]. The position of the target was randomized across trials. Along with the presentation of the pictures, a blue circle appeared at the center of the screen. After 500 ms, the circle turned red, cueing the participant to click on it to start the trial. This allowed the participants to briefly look at the

[3] That is, if the item pairs were not fixed, participants would be able to figure out the word-picture mappings during testing as a result of cross-situational learning (Yu & Smith, 2007).

pictures before hearing anything, thus minimizing eye movements due to visual search (rather than lexical processing). As soon as participants clicked on the red circle, it disappeared and an auditory stimulus was played through the headphones. Participants then clicked on the picture they believed to be the referent of the word. No feedback was provided during testing. Each testing block lasted approximately 5 minutes.

Participants completed all four blocks within the same session, and were given a chance to take a break every 16 trials (both for the training and testing Phases). There was no time limit on the trials; however, participants typically responded in less than 2 sec (M = 1,176 ms, SD = 256 ms).

Eye-tracking Recording and Analysis

Participants were calibrated using the standard 9-point display and monocular eye movements were recorded at a sampling rate of 1,000 Hz (but were resampled at 250 Hz during pre-processing, which is standard for this type of data). As in previous studies (Kapnoula & Samuel, 2019; McMurray et al., 2002), this was automatically parsed into saccades and fixations using default psychophysical parameters. Adjacent saccades and fixations were combined into a single “look” that started at the onset of the saccade and ended at the offset of the fixation.

Eye movements were recorded from the onset of the trial (presentation of unfamiliar object for training trials; red circle for testing trials) through the participant’s response (mouse click). This resulted in a variable trial offset time, depending on the individual response time. We adopted the approach of many prior studies (Allopenna et al., 1998; McMurray et al., 2002) by setting a fixed trial duration of 2,000 ms. If a trial ended before this point, we extended the last eye movement; trials longer than 2,000 ms were truncated. This approach assumes that any fixations made in the very late portions of a trial reflect the word the participant settled on and should, thus, be interpreted as an estimate of the final state of the system.

In converting the coordinates of each look to the object being fixated, the boundaries of the regions of interest containing the objects were extended by 100 pixels in order to account for noise and/or head-drift in the eye-track record. This did not result in any overlap between the objects (the dead space between pictures was 224 pixels).

Results

Two participants were excluded from the analyses of fixations (but were included in the analyses of responses) due to eye-tracking problems[4].

Analyses of responses

Training. Participants performed the task without difficulties and responded in a prompt manner. In addition, their responses were checked offline by a trained research assistant, who verified that they were doing the task as requested. Specifically, spoken responses from the production task were processed with CheckVocal (Protopapas, 2007) to check accuracy[5] and placement of response time (RT) marks. Accuracy was at 100% across participants, and average RT was 648 ms (SD = 129 ms).

Testing. Average accuracy in testing was 98.1% (SD = 4.9%), which corresponds to 1.2 error trials (out of 64) per participant. Only correct trials were included in the reaction time (RT) analyses. Average RT was 1,210 ms (SD = 267 ms).

We assessed the effects of training condition (Perception-Only versus Production) and length of training (Phase 1 versus Phase 2) on accuracy (logit-transformed) and RT using 2 × 2

[4] These participants seemed to use their peripheral vision instead of looking directly at the pictures. The same applies to the exclusion of participants due to eye-tracking problems in Experiments 2 and 3.
[5] An utterance was marked as correct only if all phonemes were pronounced correctly.

repeated measures ANOVAs. For accuracy, Condition was not significant, $F_{(1,39)}=0.241$, $p=.626$, $\eta^2=.006$, but Phase was, $F_{(1,39)}=9.144$, $p=.004$, $\eta^2=.190$, with participants showing higher accuracy in Phase 2 (99.1% compared to 97.1% in Phase 1). The interaction was not significant, $F_{(1,39)}=0.218$, $p=.643$, $\eta^2=.006$. Similarly, for RT, Condition was not significant, $F_{(1,39)}=0.097$, $p=.757$, $\eta^2=.002$, but Phase was, $F_{(1,39)}=45.952$ $p<.001$, $\eta^2=.541$, with participants giving faster responses in Phase 2 (1,092 ms compared to 1,327 ms in Phase 1). The Phase×Condition interaction was not significant, $F_{(1,39)}=0.041$, $p=.841$, $\eta^2=.001$. Overall, the behavioral results showed that participants were faster and more accurate on the Phase 2 test.

Analyses of fixations

Next, we analyzed participants' eye-movements. These analyses only include testing trials. For all analyses of eye-movements, we adopted the linking hypothesis originally proposed by Allopenna et al. (1998), according to which fixation proportions can be used as a direct measure of lexical activation.

Replication of Zamuner et al. (2016). We started by looking only at the data directly comparable to the data reported by Zamuner et al. (2016). These correspond to the testing block of Phase 1. Similarly to Zamuner et al. (2016), we opted for an analysis sensitive to the dynamic changes of lexical activation over time. However, in contrast to the original study, we fitted our data using a nonlinear curve-fitting approach (Farris-Trimble & McMurray, 2013; McMurray et al., 2010; Seedorff et al., 2018), rather than a growth curve analysis (GCA; Mirman et al., 2008). Curve-fitting, like GCA, is not restricted to a specific time window; this analytical approach is based on taking each participant's fixation data (per condition) and finding a set of parameter values that best describes the shape of the fixation curve as a whole. This means that we do not test for differences in specific time-windows, but rather differences in aspects of the overall shape of the data. Specifically, curve-fitting parameters can be mapped onto psychologically meaningful aspects of lexical activation. For example, given the linking hypothesis described above (according to which, fixation proportions map to lexical activation), the steepness of the ascending slope can be mapped to speed of activation. This transparency provides a more straight-forward comparison between experimental conditions (Farris-Trimble & McMurray, 2013; McMurray et al., 2008; Scheepers et al., 2008).

We fit our data using a four-parameter logistic function (see Eq.1. in McMurray et al., 2010). In this equation, the lower/higher asymptotes correspond to the baseline/peak of the curve (i.e., minimum and maximum lexical activation) respectively, the slope reflects how quickly lexical activation builds up in time, and the crossover corresponds to the point in time when activation crosses from the lower half of the range to the higher half (e.g., if baseline is 0 and peak is 1, then the crossover would correspond to the point in time when activation is .5).

All curve-fitting analyses were implemented using the bdots R package (Oleson, Cavanaugh, McMurray, & Brown, 2015; Seedorff, Oleson, Brown, Cavanaugh, & McMurray, 2017). First, we computed the average proportion of looks to the target for each time point along the entire time-course of the trial (i.e., 0 – 2,000 ms) separately for each subject and each training condition (see Figure 2).

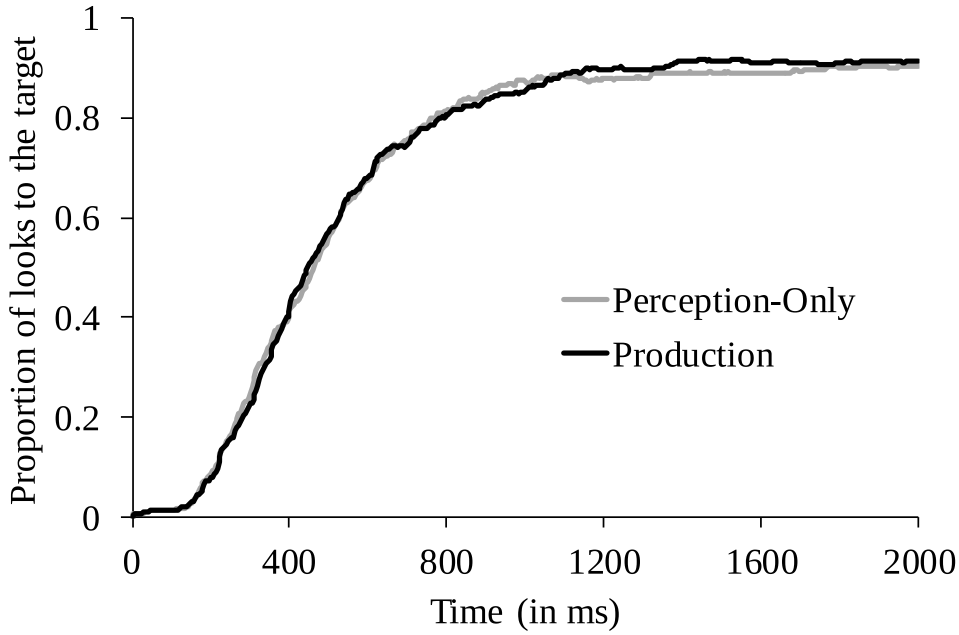

Figure 2. Proportion of looks to the target in time for each training condition (Perception-Only versus Production) in the first testing block of Experiment 1.

Then, we used the bdots logistic.fit function to find the four-parameter logistic function that provided the best fit for each curve. Using the bdots logistic.boot function (for paired data), we tested the effect of training condition on each of the three parameters of interest (peak, slope, and crossover; see Seedorff, Oleson, Cavanaugh, & McMurray, 2017; Seedorff et al., 2018, for a presentation of the bdots package, its conceptual implementation, and a discussion of the statistical approach).

Table 2. *Comparisons for the logistic parameters used to describe target activation in the testing block of Phase 1 (Experiment 1)*

Parameter	Difference between conditions (Perception-Only minus Production)	$t_{(35)}$	SE	p
Peak (*p*)	-0.0275	-7.121	0.004	<0.001
Slope (*s*)	-0.0002	-1.962	<0.001	0.058
Crossover (*c*)	-6.8892	-0.713	9.662	0.481

Note 1. All *t* tests rely on simulations; they are calculated using the bootstrapped means and are adjusted to account for the additional variance around the parameter estimate (Seedorff et al., 2017, 2018).
Note 2. Differences in dfs between phases are due to different number of excluded bad fits.
These notes also apply to Tables 3, 4, 5, 6, and 7.

As seen in Table 2, there was a significant difference between conditions in their asymptotes (reflecting maximum lexical activation). The direction of this effect indicates an advantage for the Production condition, meaning ultimately higher lexical activation for words that were produced during training. In addition, there was a difference in the slopes (reflecting speed of activation build-up) in the same direction (i.e., Production advantage), but it was not significant. These results are consistent with the findings reported by Zamuner et al. (2016).

Modulation of the production effect by length of training. Next, we asked whether this positive effect of production is maintained after additional training[6]. We thus looked at the trials from the testing block of Phase 2. Average proportions of looks to the target are plotted in Figure 3.

[6] In all three Experiments, we examined the effect of Condition as a function of training length by testing the difference between Conditions in each Phase and then comparing the results between Phases. A direct statistical comparison of the Condition effect between Phases is reported below in the **Summary and additional analyses across Experiments**; see full 3×2×2×2 (Experiment×Condition×Phase×TimeWindow) ANOVA.

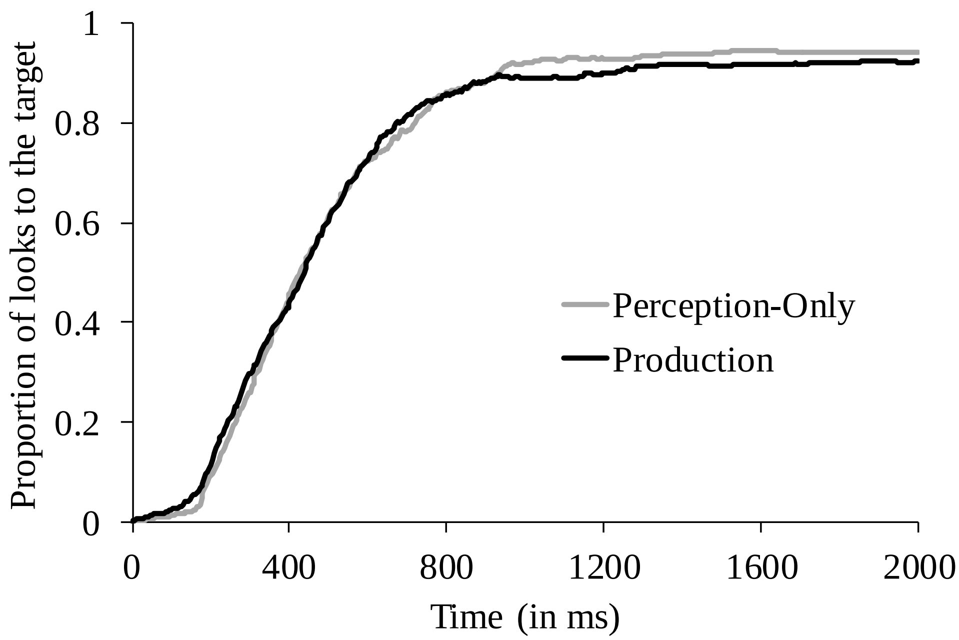

Figure 3. Proportion of looks to the target in time for each training condition (Perception-Only versus Production) in the second testing block of Experiment 1.

The same analytical approach was adopted as above.

Table 3. *Comparisons for the logistic parameters used to describe target activation in the testing block of Phase 2 (Experiment 1)*

Parameter	Difference between conditions (Perception-Only minus Production)	$t_{(34)}$	SE	p
Peak (*p*)	0.0144	3.722	0.004	<0.001
Slope (*s*)	-0.0001	-1.531	<0.001	0.135
Crossover (*c*)	2.9198	0.277	10.534	0.783

As seen in Table 3, there was again a significant difference between conditions in their asymptotes. However, in this case the direction of the effect indicated an advantage for the Perception-Only condition (i.e., ultimately higher lexical activation for words assigned to the

Perception-Only condition, or put another way, lower lexical activation for words that were produced during training).

Discussion

In line with Zamuner et al. (2016), Experiment 1 showed an early advantage for novel words that were repeated during training; they were more strongly activated after a few training trials. However, this effect was reversed after additional training; with more training, production had a detrimental effect.

One possible interpretation of this reversal is that the effect of production depends on the learning stage. That is, production may be particularly helpful during the first stages of word learning, perhaps facilitating the very first encoding of a novel word; in contrast, when it comes to integrating the lexical representation into the mental lexicon, its effect may be more harmful than helpful. This later detrimental effect could be due to 1) production being disruptive, 2) perception-only being more helpful, or 3) both (see further discussion in the General Discussion).

As discussed in the Introduction, one limitation of the Zamuner et al. (2016) design is that it confounds production with speaker variability. That is, words assigned to the production condition were also spoken by one additional voice during training. Prior work has shown a facilitatory role of talker variability in word learning (Richtsmeier et al., 2009; Rost & McMurray, 2009). Thus, any advantage for the Production condition could be due to the additional talker in training. To address this, we ran Experiment 2, which was otherwise identical to Experiment 1, but controlled for the effect of speaker variability.

Experiment 2

Experiment 2 assesses the effect of production independently of speaker variability and provides an additional test of the reversal of the effect as a function of amount of training. To match the number of talkers between conditions, we added an additional talker in the Perception-Only condition. If the facilitatory effect of the Production condition was due to talker variability, it should disappear in Experiment 2.

Method

Participants

Forty-one (28 females; mean age = 24.3 years) native speakers of Spanish participated in Experiment 2. Experiment 2 was identical to Experiment 1 in terms of participant characteristics, compensation, and ethical approval procedures.

Design

The same design as that of Experiment 1 was used with one critical difference: auditory stimuli in the Perception-Only condition were presented in two different speaker voices. As a result of this change, participants in both groups heard two different voices on each training trial (either two different voices played to them, or one played to them plus their own voice).

Materials

All items were the same as in Experiment 1. The only difference was that all items were recorded by an additional speaker of a different gender (a native male speaker of Spanish). In addition, we controlled for both inter-and intra-talker variability; to match the acoustic variability naturally present in the participants' own utterances (i.e., the fact that each time a speaker says a word the acoustics are slightly different), we selected multiple recordings of the words, such that each token was only heard once during the entire experiment (i.e., there were twelve different tokens for each word, as many as the training repetitions per word). The new recordings were pre-processed following the same steps as in Experiment 1. The average duration of the new stimuli (including the 100 ms of silence) was 694 ms. The new items were presented during training in the Perception-Only condition after the "listen" prompt (i.e., for the second presentation of each word). The original stimuli from Experiment 1 were used in training for the first presentation (i.e., before the visual prompt) for both conditions, and during testing.

Visual stimuli were the same as in Experiment 1.

Procedure

The procedure was identical to that of Experiment 1.

Eye-tracking Recording and Analysis

Eye-tracking recording and pre-processing were identical to those of Experiment 1.

Results

Analyses of responses

Training. Participants performed the task without problems and their responses were checked offline by a trained research assistant, who verified that they were doing the task as requested. That is, as in Experiment 1, spoken responses from the production task were processed with CheckVocal (Protopapas, 2007). Accuracy was at 100% across participants, while average RT was 705 ms (SD = 128 ms).

Testing. Average accuracy in testing was 96.4% (SD = 5.3%), which corresponds to 2.3 error trials (out of 64) per participant. Any trials with incorrect responses were excluded from RT analyses. In the remaining trials, average RT was 1,320 ms (SD = 271 ms).

We assessed the effects of training condition and length of training on accuracy (logit-transformed) and RT using the same analytical approach as in Experiment 1. For accuracy, Condition was not significant, $F_{(1,40)}$=0.291, p=.592, η^2=.007, but Phase was, $F_{(1,40)}$=16.956, p<.001, η^2=.298, with participants showing higher accuracy in Phase 2 (99.2% compared to 93.7% in Phase 1). The interaction was not significant, $F_{(1,40)}$=1.321, p=.257, η^2=.032. For RT, Condition was significant, $F_{(1,40)}$=6.186, p=.017, η^2=.134, with participants giving faster responses for items that had been in the Perception-Only condition during training (1,284 ms compared to 1,371 ms for Production items). Phase was also significant, $F_{(1,40)}$=31.220, p<.001, η^2=.438, with participants giving faster responses in Phase 2 (1,171 ms compared to 1,484 ms in Phase 1). The interaction did not reach significance, $F_{(1,40)}$=2.849, p=.099, η^2=.066. Even though the interaction was not significant, in the interest of comprehensiveness, we conducted Bonferroni-corrected post-hoc comparisons, which revealed a detrimental effect of Production in Phase 1, $F_{(1,40)}$=4.830, p=.034, η^2=.108, which was in the same direction, but not significant, in Phase 2, $F_{(1,40)}$=3.617, p=.064, η^2=.083.

Overall, the behavioral results showed that participants were faster and more accurate in Phase 2, but they were also consistently faster in recognizing items that had been assigned to the Perception-Only training condition.

Analyses of fixations

Next, we analyzed participants' eye-movements during testing.

Modulation of production effect by length of training. We first looked at the effect of production as a function of training length (i.e., Phase 1 versus Phase 2; average proportions of looks to the target in time for each Phase are plotted in Figures 4 and 5). The same analytical approach was adopted as in Experiment 1 (i.e., curve-fitting).

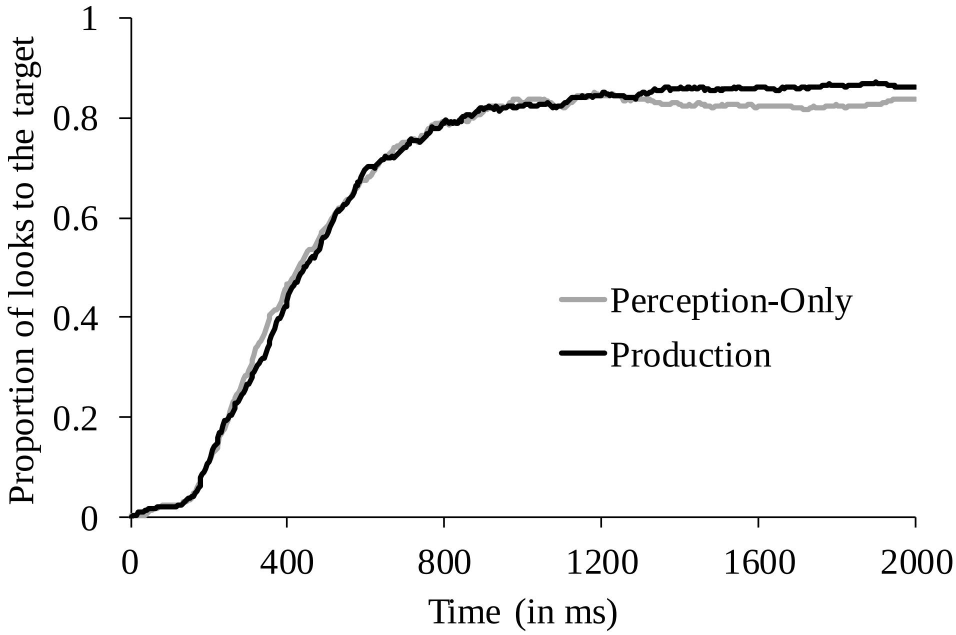

Figure 4. Proportion of looks to the target in time for each training condition (Perception-Only versus Production) in the first testing block of Experiment 2.

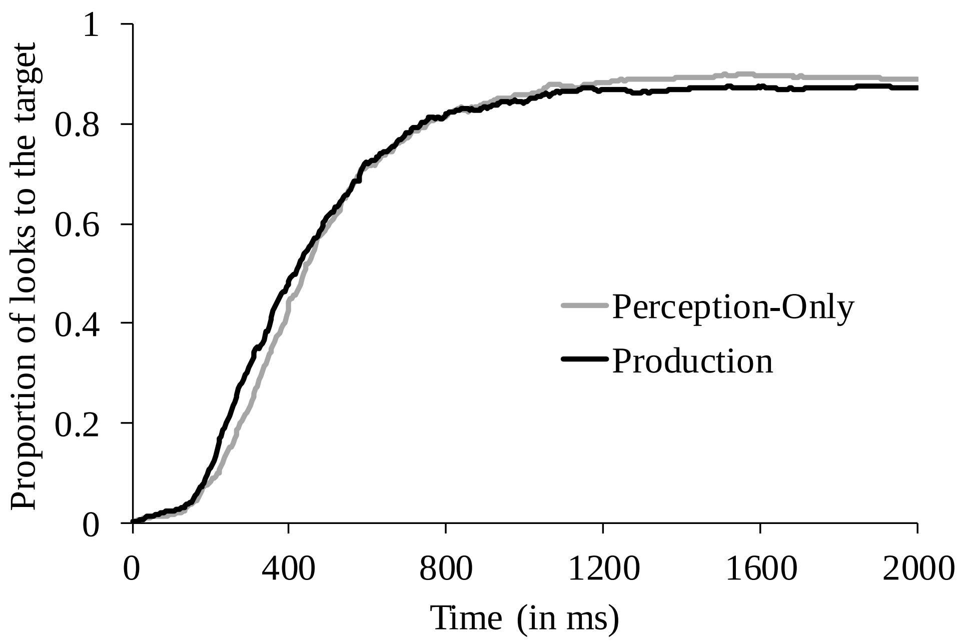

Figure 5. Proportion of looks to the target in time for each training condition (Perception-Only versus Production) in the second testing block of Experiment 2.

Table 4. *Comparisons for the logistic parameters used to describe target activation in the testing block of Phase 1 (Experiment 2)*

Parameter	Difference between conditions (Perception-Only minus Production)	$t_{(32)}$	SE	p
Peak (*p*)	-0.0188	-2.749	0.007	0.001
Slope (*s*)	0.0010	8.644	<0.001	<0.001
Crossover (*c*)	-12.2863	-0.542	22.674	0.592

Table 5. *Comparisons for the logistic parameters used to describe target activation in the testing block of Phase 2 (Experiment 2)*

Parameter	Difference between conditions (Perception-Only minus Production)	$t_{(36)}$	SE	p
Peak (*p*)	0.0090	2.714	0.003	0.010
Slope (*s*)	-0.0001	-1.197	<0.001	0.239
Crossover (*c*)	53.188	5.546	9.256	<0.001

The results were similar to Experiment 1 in two respects. First, there was an early facilitatory effect of production (significantly higher activation peak for produced items after the first training Phase; see Table 4). Second, after the second training Phase, the efficacy of the two training conditions reversed, resulting in higher lexical activation for words assigned to the Perception-Only condition (see Table 5).

In contrast to Experiment 1, there was a significant difference in slope in the first testing Phase, indicating faster activation for Perceived-Only items, and a significant difference in crossover in the second testing phase, indicating an earlier onset of activation for produced items (see Discussion below). A direct comparison across all three experiments is presented in the **Summary and additional analyses across Experiments** section below.

Discussion

The critical difference between Experiments 1 and 2 was that the latter controlled for the effect of speaker variability. This allowed us to better examine the effect of production in and of itself. Experiment 2 replicated the reversal of the production effect that was observed in Experiment 1: an early facilitatory effect of production turned into a detrimental effect after additional training (i.e., a significantly higher activation peak for Perceived-Only items in Phase 2).

In Experiment 2, we found a steeper slope for items assigned to the Perception-Only condition (i.e., faster activation). Even though this effect may seem inconsistent with the early facilitatory effect of production (i.e., higher activation asymptote for produced items), it could reflect the way in which novel items are gradually integrated into the mental lexicon. For example, one possibility is that production facilitates early encoding of novel lexical representations (the "lexical configuration" stage proposed by Leach & Samuel, 2007), but uninterrupted perception leads to better overall integration into the mental lexicon (the "lexical engagement" stage proposed by Leach & Samuel, 2007). The automatization of lexical processing is considered to be a marker of deeper lexical integration (i.e., full lexical engagement; see discussion by McMurray, Kapnoula, and Gaskell, 2016). From this perspective, the slope should in fact be steeper for Perception-Only items to the extent that it reflects automatization of processing.

In Phase 2, our analysis showed a later crossover for items assigned to the Perception-Only condition. At first, this seems to be a surprising result (given that in the same Phase we observed overall higher activation of Perception-Only items), but it is in fact consistent with the rationale laid down in the previous paragraph, according to which production may facilitate early encoding (lexical consolidation), but perception may lead to better overall integration (lexical engagement). In other words, the *onset* of activation (better reflected by the crossover parameter) may rely on the configuration status of a novel word, whereas the *speed* of activation (better reflected by the slope parameter) should depend on the degree of automatization. Thus, if production facilitates lexical configuration, it would make sense to see a facilitatory effect of production on crossover.

Taken together, the results from Experiments 1 and 2 suggest that producing a novel word may have an early advantage, but its effect becomes detrimental with additional training. In Experiment 3, we consider the possibility that production may play a positive role by promoting the abstraction of newly acquired lexical representations, in which case we should observe better generalization of learning to novel instances of learned words. In Experiments 1 and 2 the testing voice was the same as at least one of the voices used in training. That means that during testing all items (in both conditions) were heard in a familiar voice, one that had been heard before. Our new question is whether Production (versus Perception-Only) might help learners *generalize* to novel speakers. To address this question, in Experiment 3, all testing items were presented in a new voice. If Production helps generalization, we should see a stronger facilitatory effect of Production in Phase 1 and perhaps a weaker detrimental effect in Phase 2. In contrast, if Perception-Only leads to better generalization via stronger lexical engagement, we should see the opposite pattern.

Experiment 3

Experiment 3 tests whether production helps listeners in recognizing novel words spoken by an unfamiliar talker (i.e., generalization). To assess this, test stimuli were presented in a novel voice. As in Experiment 2, we matched the number of talkers between training conditions by having an additional talker in the Perception-Only condition.

Method

Participants

Forty-one (28 females; mean age = 24.9 years) native speakers of Spanish participated in Experiment 3. Experiment 3 was identical to Experiments 1 and 2 in terms of participant characteristics, compensation, and ethical approval procedures.

Design

The same design as that of Experiment 2 was used with one critical difference: auditory stimuli presented in testing were spoken by one of two new speakers (one male and one female). Twenty participants were randomly assigned to one speaker and the rest to the other.

Materials

All training stimuli were identical to those of Experiment 2. The only difference between Experiments 2 and 3 was in the testing stimuli. All items were recorded by two additional native speakers of Spanish (one male and one female). These items were used to replace the testing stimuli used in Experiments 1 and 2. As with the male training items of Experiment 2, we selected multiple recordings of the words, such that each token was only heard once during the entire experiment (i.e., there were eight different tokens for each word). This was done to increase acoustic variability of the testing stimuli, as in Experiment 2. The new recordings were pre-processed following the same steps as in Experiments 1 and 2. The average duration of the new stimuli (including the 100 ms of silence) was 624 ms for the male and 840 ms for the female speaker.

Visual stimuli were the same as in Experiment 1.

Procedure

The procedure was identical to that of Experiments 1 and 2.

Eye-tracking Recording and Analysis

Eye-tracking recording and pre-processing were identical to that of Experiments 1 and 2.

Results

Two participants were excluded from the analyses of fixations (but were included in the analyses of responses) due to eye-tracking problems.

Analyses of responses

Training. Participants performed the task without problems and their responses were checked offline by a trained research assistant, who verified that they were doing the task as requested. Once again, spoken responses from the production task were processed with CheckVocal (Protopapas, 2007). Accuracy was at 100% across participants, while average RT was 689 ms (SD = 144 ms).

Testing. Average accuracy in testing was 97.3% (SD = 4.5%), which corresponds to 1.7 error trials (out of 64) per participant. Any trials with incorrect responses were excluded from RT analyses. In the remaining trials, average RT was 1,256 ms (SD = 227 ms).

We assessed the effects of training condition and length of training on accuracy (logit-transformed) and RT using the same analytical approach as in Experiments 1 and 2. Accuracy was higher for Production (97.9%) compared to Perception-Only items (96.9%), but this difference was not significant, $F_{(1,40)}$=3.121, p=.085, η^2=.072. Phase was significant, $F_{(1,40)}$=21.168, p<.001, η^2=.346, with higher accuracy in Phase 2 (99.7% compared to 95.0% in Phase 1). The interaction was not significant, $F_{(1,40)}$=0.205, p=.653, η^2=.005. For RT, Condition was not significant, $F_{(1,40)}$=0.710, p=.404, η^2=.017, but Phase was, $F_{(1,40)}$=49.961, p<.001, η^2=.555, with participants giving faster responses in Phase 2 (1,125 ms compared to 1,398 ms in Phase 1). The interaction was not significant, $F_{(1,40)}$=1.378, p=.247, η^2=.033.

Overall, the behavioral results showed that participants were faster and more accurate in Phase 2.

Analyses of fixations

Next, we analyzed participants' eye-movements during testing.

Modulation of production effect by length of training. The same analytical approach was adopted as in Experiments 1 and 2 (i.e., curve-fitting). Average proportions of looks to the target in time for each Phase are plotted in Figures 6 and 7.

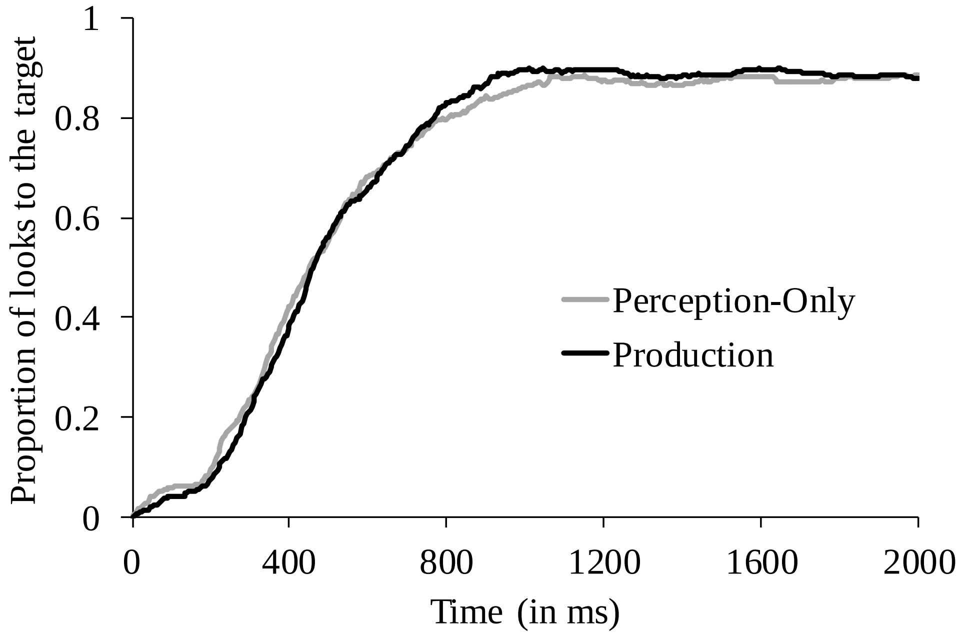

Figure 6. Proportion of looks to the target in time for each training condition (Perception-Only versus Production) in the first testing block of Experiment 3.

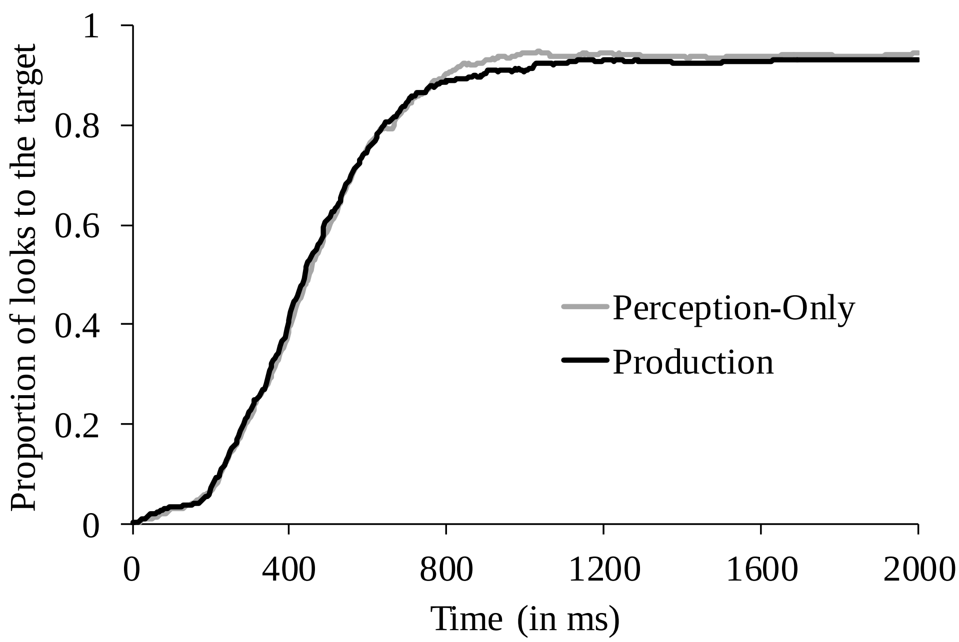

Figure 7. Proportion of looks to the target in time for each training condition (Perception-Only versus Production) in the second testing block of Experiment 3.

Table 6. *Comparisons for the logistic parameters used to describe target activation in the testing block of Phase 1 (Experiment 3)*

Parameter	Difference between conditions (Perception-Only minus Production)	$t_{(34)}$	SE	p
Peak (*p*)	-0.0063	-0.978	0.007	0.335
Slope (*s*)	0.0001	0.747	<0.001	0.460
Crossover (*c*)	0.369	0.044	10.444	0.972

Table 7. *Comparisons for the logistic parameters used to describe target activation in the testing block of Phase 2 (Experiment 3)*

Parameter	Difference between conditions (Perception-Only minus Production)	$t_{(37)}$	SE	p
Peak (*p*)	0.0115	3.117	0.004	0.004
Slope (*s*)	-0.0001	-1.142	<0.001	0.261
Crossover (*c*)	21.6634	3.937	5.502	<0.001

In contrast to Experiments 1 and 2, there was no early facilitatory effect of production (see Table 6). However, in line with the previous experiments, we again observed higher lexical activation for words assigned to the Perception-Only condition in the testing block of Phase 2 (see Table 7). In addition, similarly to Experiment 2, there was again a significant crossover difference in Phase 2, indicating an earlier activation onset for produced items (see Table 7).

Discussion

As in Experiments 1 and 2, we again observed a detrimental effect of production after additional training (i.e., significantly higher activation peak for Perception-Only items in Phase 2). In addition, as in Experiments 1 and 2, the activation peak was higher for Production compared to Perception-Only items in Phase 1, though in this case the effect was not significant.

As in Experiment 2, we again found an earlier crossover for produced items in Phase 2. This effect is in line with the idea (see the discussion in Experiment 2) that the onset of lexical activation (as reflected by the crossover parameter) depends on the early configuration of a novel word, rather than its integration into the lexicon.

Summary and additional analyses across Experiments

Experiments 1 and 2 replicated the finding reported by Zamuner et al. (2016) that production helps at the early stages of word learning. In Experiment 3, the results from Phase 1 were in the same direction, but were not significant. In contrast, and critically, we observed a reliable detrimental effect of production across all three experiments after additional training (i.e., higher activation of non-produced words in Phase 2). This reversal of the effect seems to reflect a dissociation regarding the time course of lexical integration: *Even though early encoding may be facilitated by production, further integration is better served by perception*. In line with this interpretation, we also observed a facilitatory effect of production on the crossover (likely reflecting earlier activation onset), which contrasted with a facilitatory effect of perception on the slope (likely reflecting automatization of processing). That is, the facilitation of production on lexical encoding may be reflected by the earlier activation onset (crossover), whereas the facilitation of perception on lexical integration may be reflected in the activation speed (slope).

We conducted additional analyses to examine these patterns in greater detail. Specifically, we were interested in testing for any significant differences between early and late stages of lexical processing. To do so, we split each trial into two (early/late) parts. Since this was a post-hoc, exploratory analysis, we chose to avoid splitting the trials based on an arbitrary, experimenter-driven criterion. Instead, each trial was split based on the offset of the auditory stimulus (corrected for 200 ms oculomotor delay). This allowed us to use a flexible, stimulus-driven time window and account for any variability between experiments, speakers, and stimuli[7]. In addition, this analysis allowed us to directly test all possible interactions between independent variables.

Results

Here we focused on the effect of training condition (Perception-Only versus Production) on the activation of novel words as a function of 1) amount of training and 2) time-point within a trial. Results from all three experiments were included in the analyses (see Figure 8). We started by running the full 3 (Experiment: 1/2/3) × 2 (Condition: Perception-Only/Production) × 2 (Phase: 1/2) × 2 (TimeWindow: early/late) repeated-measures ANOVA with average proportion of fixations to the target (empirical-logit-transformed) as the DV. Detailed results of this ANOVA and its follow-ups are listed in the Appendix.

Phase was significant, $F_{(1,114)}$=45.665, p<.001, η^2=.286, indicating that participants were better at activating the target word after receiving additional training, as expected. Condition was not significant, $F_{(1,114)}$=0.013, p=.909, η^2<.001, and neither was the Condition × Phase interaction, $F_{(1,114)}$=3.107, p=.081, η^2=.027. The Condition × Phase × TimeWindow interaction was significant, $F_{(2,114)}$=13.481, p<.001, η^2=.106, reflecting a differential effect of training condition depending on training length and time-point within trial. Neither the 4-way, nor any of the other 3-way interactions were significant.

[7] A flexible splitting point that is time-locked to each stimulus takes into account the variability in stimulus duration. This leads to time-windows that are informationally comparable between items. Furthermore, a splitting point roughly close to the middle of the trial (as is the case with the splitting point used) means that the two time windows are comparable in terms of number of data points.

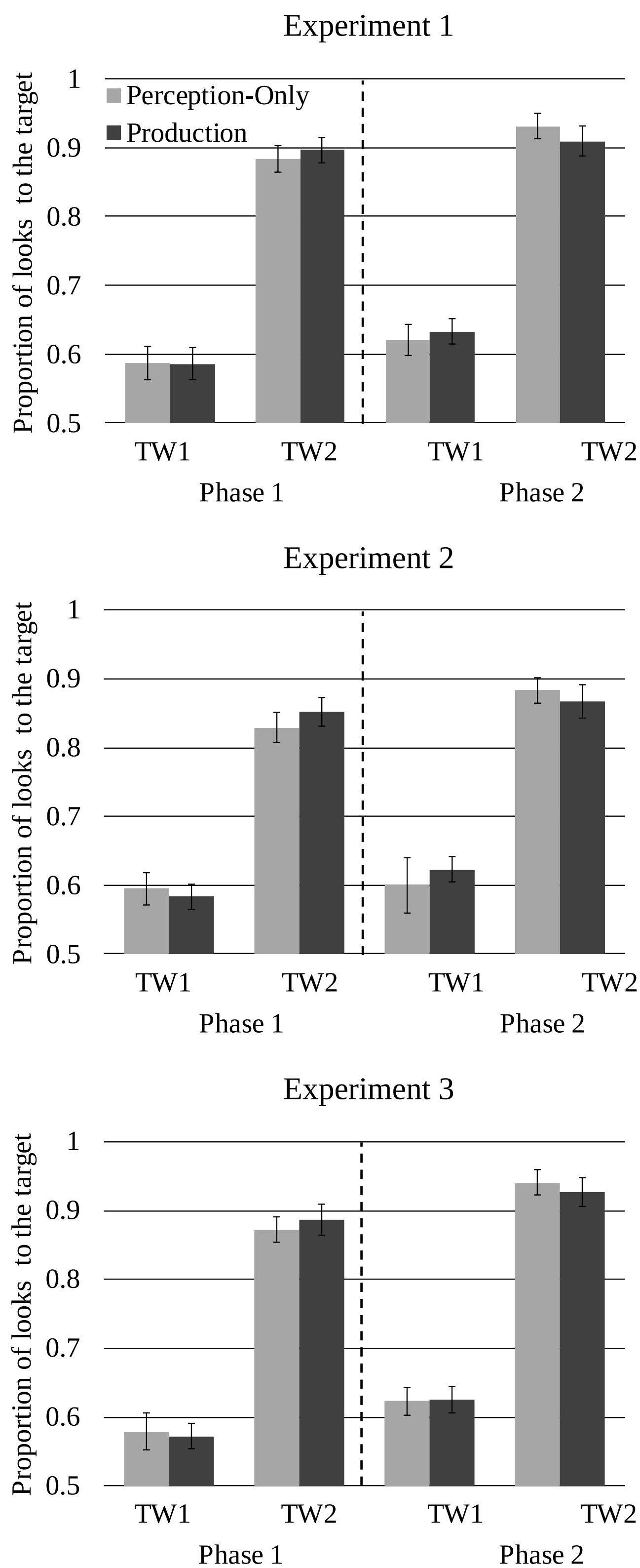

Figure 8. Average proportions of looks to the target per training condition (Perception-Only/Production), testing phase (Phase 1/Phase 2), and time window (TW) within trial (early: TW1/late: TW2) for each of the three experiments. Error bars indicate ±1within-subject standard

error of the mean (Cousineau, 2005; Loftus & Masson, 1994; Morey, 2008). *Note*. An alternative visualization of the data is offered in Figure A1 in the Appendix.

In following up on the significant Condition × Phase × TimeWindow interaction, post-hoc comparisons (Bonferroni-corrected) showed a significant Condition effect in the late TimeWindow of Phase 2, p=.022, the direction of which indicates more looks to the Perception-Only items late in the trial. In contrast, in the early TimeWindow of Phase 2, the trend was in the opposite direction (i.e., more looks to Production items; p=.059). No simple effect of Production was found in Phase 1.

The results of these within-trial analyses show a shift in the direction of the Production effect in Phase 2; specifically, if there is any effect of production early in a trial, this effect seems to be facilitatory; later in the trial, production is clearly detrimental. This within-trial pattern echoes the curve-fitting results: Even though production seems to expedite the activation onset of new words, providing an advantage to produced items early in the trial, items that had only been perceived, without production, enjoy a faster build-up of activation.

General Discussion

Across three experiments, we examined the role of production on word learning while carefully controlling for other variables such as amount of exposure, talker variability, recall-based facilitation, and attention. Our findings show a robust dissociation: At first production helps, but as learning advances its effect becomes detrimental. This dissociation can potentially reconcile previous results showing both facilitatory (Dodson & Schacter, 2001; Gathercole & Conway, 1988; Hopkins & Edwards, 1972; Hopman & MacDonald, 2018; P. MacDonald & MacLeod, 1998; MacLeod et al., 2010; MacLeod & Bodner, 2017; Zamuner et al., 2016) and detrimental (Baese-Berk, 2019; Baese-Berk & Samuel, under review, 2016; Leach & Samuel, 2007; Zamuner et al., 2018) effects of production. Moreover, our findings provide fine-grained timing information about the effect of production on learning – both at the level of training phase (i.e., few versus many training trials) and in terms of real-time processing, at the single-trial scale. In that sense, the present work complements previous work such as the Leach & Samuel's (2007) study that showed larger negative effects of production on perceptual learning but without offering timing information. As a result, our results offer valuable insights into the mechanisms underlying the seemingly contrasting effects of production.

Early facilitatory effect of production

We found evidence for an early facilitatory effect of production across three experiments. In Experiments 1 and 2, novel words that had been produced during training had a significantly higher activation peak after the first training block (i.e., an effect on the peak); Experiment 3 showed the same pattern, but the difference did not reach significance. In other words, production was helpful during the earliest stages of word learning, i.e., during the initial encoding of novel word-forms.

In addition, in Experiments 2 and 3, produced items had an advantage at the onset of lexical activation (i.e., an effect on the crossover). The early advantage for production was echoed in our within-trial analyses (the significant Condition × Phase × TimeWindow interaction). We suggest that this early activation advantage is a result of production helping the *initial* encoding of novel words. Better lexical configuration of a novel word-form can, in turn, facilitate an earlier onset of activation. Note that in Leach and Samuel's (2007) original contrast between lexical configuration and lexical engagement, an important basis for the distinction was their finding of a facilitatory effect of production on lexical configuration, consistent with this finding.

Late detrimental effect of production

Evidence for a detrimental effect of production was even more robust; across all experiments and analyses, produced words had a lower activation peak after the second training block. In addition, produced words displayed slower activation build-up (i.e., an effect on the slope) in Experiment 2. This pattern likely indicates that words that were only heard (not produced) were better integrated into the mental lexicon. That is, once a word is well integrated into the system, its recognition reaches a higher level of automatization. As a result, the recognition process moves faster (reflected by the higher slope) and is more effective (reflected by the higher peak).

This finding conforms to Leach and Samuel's (2007) concept of lexical engagement, with their results showing better lexical engagement for listeners in a Perception-Only training condition than in a Production condition. This interpretation is also in line with Kapnoula and Samuel (2019), who found that newly learned words were activated faster after participants slept (indicated by the steeper slopes of fixation probability curves). Given the well-documented strengthening role of sleep consolidation in lexical integration, this can be taken as indirect evidence that better integrated words show a more robust (faster and/or higher) pattern of activation.

Towards a reconciling mechanism

As discussed in the Introduction, adding a production requirement may affect the outcome of word learning in a number of different ways, some of which are not intrinsic to production per se. Our study was designed to control for confounding factors as much as possible, which potentially allows us to identify the mechanism(s) driving any true production effect. For example, participants were always forced to keep the new phonological sequences in their phonological short-term memory until they saw the prompt, thus equalizing the role of such memory effects across conditions. The same aspect of the procedure also controlled for any differences in attention. In addition, no recall was required during training (only immediate repetition), meaning that any production effect could not be driven by the testing effect (Karpicke & Roediger, 2008). Lastly, in Experiments 2 and 3, we controlled for the effect of speaker variability, which is another frequently confounding variable.

With these potentially confounding factors controlled, our experiments were designed to examine *dynamic* effects of producing to-be-learned words. That is, we tested how production affects lexical activation in real time using the high temporal resolution of eye tracking, and how this effect may change during the progression of learning.

Our results show an early facilitatory effect of production followed by a late detrimental effect. This pattern can reconcile a number of previously reported findings. As discussed earlier, the pattern aligns very well with the theoretical dissociation proposed by Leach and Samuel (2007), according to which lexical configuration precedes lexical engagement. From this perspective, production seems to help early lexical encoding (configuration), but it hurts lexical integration (engagement). Consistent with previous findings (Kapnoula et al., 2015; Kapnoula & McMurray, 2016), we found evidence for both lexical properties being developed within minutes[8] after the onset of learning. This indicates that any internal adjustments that are made to support the development of these two properties (e.g., formation of bottom-up, lateral, and top-down pathways connecting the novel words to other, known representations) may be

[8] In Kapnoula and McMurray (2016) and Kapnoula et al. (2015), evidence for integration was found within 15-30 mins after learning onset.

theoretically and mechanistically distinct, but their development unfolds in a cascaded and possibly overlapping fashion.

Most importantly, our findings and their interpretation within this dual-stage theoretical framework, bring us closer to a comprehensive mechanism of the production effect. For example, as mentioned in the Introduction, one way that production may help word learning is via the addition of articulatory information. Given our results, adding articulatory information should only help with the early encoding of new word-forms; when the learner has very little other information available, every cue can help. This pattern is reminiscent of what has been found in studies looking for motor area activation during speech perception: This can be found, but this is strongly associated with very challenging listening conditions, when other cues are much less accessible (e.g., Nuttall et al., 2016). Our results speak directly to the issue of how/when production may disrupt word learning. There is not a disruptive effect of production on the early encoding of novel word-forms. When production is disruptive, this is associated with later stages, related to lexical integration (e.g., mapping the word-form to its semantic referent).

Limitations and further questions

Although our results provide incisive information about the learning stage in which production is most disruptive, they do not speak to an additional important question: Is the detrimental effect of production in fact a true negative effect, or is the difference between training conditions (here, and in other studies) due to a facilitatory effect of perception that is reduced under production? For example, it may be that production is disruptive in the sense that participants allocate attentional resources towards producing the word and, as a result, have fewer resources available to take advantage of the perceptual input. This would be a “lost opportunity" effect, rather than an active disruption, caused by production. Indeed, this would be in line with findings showing that production becomes detrimental when cognitive load is high – e.g., because the new words are phonologically unfamiliar (Kaushanskaya & Yoo, 2011); or because they are spoken in an unfamiliar accent (Cho & Feldman, 2016); or because the resources themselves are limited due to the participants' young age (López Assef et al., 2021; Zamuner et al., 2018).

This idea is also in line with Kapnoula et al. (2015), who report similar degrees of lexical integration of new words independently of whether they were repeated during training or not. That is, production did not seem to affect lexical integration. Furthermore, Baese-Berk and Samuel (2016; under review) have examined this in the domain of learning a new phonetic contrast, and the data are consistent with this (i.e., the idea that some of the detrimental production effects are due to reduced passive exposure). If this is the mechanism, then presumably the disruption could be alleviated if the production requirement were delayed enough for the perceptual processing to finish. Recent work in our laboratory provides evidence that the same pattern holds for learning new words (Kapnoula & Samuel, under review).

Another lingering question is why would production help lexical encoding, but hurt further integration? In the Introduction, we present a number of ways in which production may help *or* hurt word learning; however, we did not expect to find evidence for both. That is, our experiments were designed to examine whether production has a positive/negative effect on different aspects and stages of word learning, but they cannot address why production has opposite effects. We speculate that these effects are driven by different mechanisms. A detrimental effect of production on lexical integration could be due to a lost opportunity for additional passive exposure (as argued above); a facilitatory effect on lexical encoding could be due to small differences in attention. Our design minimized differences in attention: participants did not know ahead of time whether they would be asked to repeat the word or not. However, it is conceivable that participants attended more to the phonological structure of a new word once they were asked to repeat it. Such differences in attention could facilitate lexical encoding. Further work is needed to examine the underlying mechanism(s) behind this complex pattern of findings.

A potential criticism of this work is that, if ultimately word recognition is highly successful, arguing for a detrimental effect of production may be a misnomer. Indeed, our results do not address the question of whether production affects the probability of correctly recognizing a newly learned word. However, our aim was to examine the effect of production on the quality of the newly learned lexical representations, as reflected by their real-time activation trajectory. In that respect, our results show that production is detrimental when compared to just hearing the new word. Here, we should note that spoken language comprehension likely depends not only on the accuracy of lexical activation, but also on its speed. According to current theories of spoken language comprehension, sentence comprehension is largely based on activation of lexical representations (Altmann & Kamide, 1999; M. C. MacDonald et al., 1994; McRae et al., 1998; Tanenhaus & Trueswell, 1995; Trueswell, 1996). More critically, there is evidence that spoken word recognition happens in parallel to semantic and syntactic processing (Gussow et al., 2019; Yee & Sedivy, 2006). This means that any delays in activating lexical representations can have important downstream consequences at higher levels of processing. Moreover, given the speed with which the speech signal unfolds, it is reasonable to assume that any delays can gradually accumulate in time, which may lead to a growing difficulty in integrating upcoming input. Thus, speed of lexical activation is likely a critical aspect of efficient spoken language comprehension.

More broadly, one may wonder about the degree to which our results are relevant to different word learning situations (e.g., in L1 versus L2, or in naturalistic versus classroom-type settings). There are indeed two aspects of our design that perhaps make the task more similar to L1 word learning; first, the novel words were phonologically, phonotactically, and morphologically consistent with the participants' L1 (Spanish); second, we used unfamiliar objects as visual referents. That said, we believe our results capture something fundamental about the cognitive mechanisms of word learning and, in that sense, our findings are relevant to and have implications for word learning in general. Somewhat related to this, even though our results do not directly speak to the question of how production can be best incorporated into word-learning practices in the real world, our findings can certainly be used as a base on which to formulate experimental hypotheses that are more directly relevant to real-world settings. For example, within a second language learning setting, one may predict that delaying the requirement for students to repeat a new word until after they have been exposed to it a few times may lead to more robust learning. Future research in more naturalistic settings can help us test such predictions.

Finally, this work examined the role of production on word learning by assessing participants' ability to recognize, rather than *produce* novel words. Our goal was to report results that are directly comparable to previous work, which also used measures of comprehension. In addition, our experimental paradigm (VWP) allowed us to detect fine differences in the dynamic build-up of lexical activation. This would have been quite difficult if we had used production at test. That said, we acknowledge that knowing a word is not limited to recognizing it.

Conclusion and significance

Our findings demonstrate an early facilitatory effect of production, followed by a (more robust) late detrimental effect. Our interpretation of this reversal is that production may facilitate early encoding of new words, but perception is more helpful when it comes to their deeper integration into the mental lexicon.

These findings are consistent with a literature that includes both facilitatory (Dodson & Schacter, 2001; Gathercole & Conway, 1988; Hopkins & Edwards, 1972; Hopman & MacDonald, 2018; P. MacDonald & MacLeod, 1998; MacLeod et al., 2010; MacLeod & Bodner, 2017; Zamuner et al., 2016) and detrimental (Baese-Berk, 2019; Baese-Berk & Samuel, under review, 2016; Leach & Samuel, 2007; Zamuner et al., 2018) effects of production, allowing us to suggest a reconciliation of the seemingly contradictory pattern of results. The theoretical framework of our interpretation is based on the two-stage dissociation proposed by Leach and Samuel (2007), according to which different lexical properties (i.e., configuration and engagement) can follow distinct developmental trajectories, shaped by different variables.

The results of our three experiments shed light on the journey of novel words into the mental lexicon. As this line of research develops, it has the potential to inform the educational community, clarifying how and when production can be used most effectively to aid novel word learning, and when it should be avoided.

Acknowledgements

The authors thank Alexandra Kypta-Vivanco for help with checking participants' responses. Support for this project was provided by the Spanish Ministry of Science and Innovation, Grant # PSI2017-82563-P, awarded to AGS and by the Spanish Ministry of Economy and

Competitiveness through the Juan de la Cierva-Formación fellowship, # FJCI-2016- 28019, awarded to ECK. This work was partially supported by the Basque Government through the BERC 2018-2021 and BERC 2022-2025 programs, and by the Spanish State Research Agency through BCBL Severo Ochoa excellence accreditation SEV-2015-0490 and CEX2020-001010-S. This project has received funding from the European Union's Horizon 2020 research and innovation program, under the Marie Skłodowska-Curie grant agreement No 793919, awarded to ECK.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language, 38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264. https://doi.org/10.1016/S0010-0277(99)00059-1

Baddeley, A. D., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. *Journal of Memory and Language*, *27*(5), 586–595. https://doi.org/10.1016/0749-596X(88)90028-9

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics*, 1–25. https://doi.org/10.3758/s13414-019-01725-4

Baese-Berk, M. M., & Samuel, A. G. (under review). *Just give it time: Differential effects of disruption and delay on perceptual learning*.

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

Boiteau, T. W., Malone, P. S., Peters, S. A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology*.

General, *143*(1), 295. https://doi.org/10.1037/A0031858

Bradlow, A. R., & Hayes, E. (2003). Speaking Clearly for Children With Learning Disabilities. *Article in Journal of Speech Language and Hearing Research*, *46*(1), 80–97. https://doi.org/10.1044/1092-4388(2003/007)

Bradlow, A. R., Torretta, G., & Pisoni, D. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*.

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

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson’s method. *Researchgate.Net*. https://doi.org/10.20982/tqmp.01.1.p042

Creel, S. C., & Tumlin, M. (2011). On-line acoustic and semantic interpretation of talker information. *Journal of Memory and Language*, *65*(3), 264–285.

Dodson, C. S., & Schacter, D. L. (2001). “If I had said it I would have remembered it”: Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin and Review*, *8*(1), 155–161. https://doi.org/10.3758/BF03196152

Duff, P. (2000). Repetition in foreign language classroom. In *The development of second and foreign language learning through classroom interaction* (pp. 109–138). Lawrence Erlbaum.

Farris-Trimble, A., & McMurray, B. (2013). Test–Retest Reliability of Eye Tracking in the Visual World Paradigm for the Study of Real-Time Spoken Word Recognition. *Journal of*

Speech, Language, and Hearing Research, *56*(4), 1328–1345.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fernandes, T., Kolinsky, R., & Ventura, P. (2009). The metamorphosis of the statistical segmentation output: Lexicalization during artificial language learning. *Cognition*.

François, C., Cunillera, T., Garcia, E., Laine, M., & Rodriguez-Fornells, A. (2017). Neurophysiological evidence for the interplay of speech segmentation and word-referent mapping during novel word learning. *Neuropsychologia*, *98*, 56–67. https://doi.org/10.1016/j.neuropsychologia.2016.10.006

Gaskell, M. G., & Dumay, N. (2003). Lexical competition and the acquisition of novel words. *Cognition*, *89*(2), 105–132. https://doi.org/10.1016/S0010-0277(03)00070-2

Gathercole, S. E. (2006). Nonword repetition and word learning: The nature of the relationship. *Applied Psycholinguistics*, *27*, 513–543. https://doi.org/10.1017.S0142716406060383

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110–119. https://doi.org/10.3758/BF03213478

Graf Estes, K., Evans, J. L., Alibali, M. W., & Saffran, J. R. (2007). Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*, *18*(3), 254–260. https://doi.org/10.1111/j.1467-9280.2007.01885.x

Gupta, P. (2003). Examining the relationship between word learning, nonword repetition, and immediate serial recall in adults. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology*, *56 A*(7), 1213–1236. https://doi.org/10.1080/02724980343000071

Gupta, P. (2008). A computational model of nonword repetition, immediate... - Google Scholar. In A. Thorn & M. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 120–147). Psychology Press.

Gupta, P., & Tisdale, J. (2009). Word learning, phonological short-term memory, phonotactic probability and long-term memory: towards an integrated framework. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1536), 3755–3771. https://doi.org/10.1098/rstb.2009.0132

Gussow, A. E., Kapnoula, E. C., & Molinaro, N. (2019). Any leftovers from a discarded prediction? Evidence from eye-movements during sentence comprehension. *Language, Cognition and Neuroscience*, 1–18. https://doi.org/10.1080/23273798.2019.1617887

Hay, J., Pelucchi, B., Estes, K., & Saffran, J. (2011). Linking sounds to meanings: Infant statistical learning in a natural language. *Cognitive Psychology*, *63*(2), 93–106.

Hebb, D. (1961). Distinctive features of learning in the higher animal. *Brain Mechanisms and Learning*, 37–46.

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews*

Neuroscience, *13*(2), 135–145. https://doi.org/10.1038/nrn3158

Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language, Cognition and Neuroscience*, *29*(1), 2–20. https://doi.org/10.1080/01690965.2013.834370

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, *69*(3), 407–422. https://doi.org/10.1016/j.neuron.2011.01.019

Höhle, B., Fritzsche, T., Meß, K., Philipp, M., & Gafos, A. (2020). Only the right noise? Effects of phonetic and visual input variability on 14-month-olds' minimal pair word learning. *Developmental Science*, *23*(5). https://doi.org/10.1111/desc.12950

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

Hopman, E. W. M., & MacDonald, M. C. (2018). Production Practice During Language Learning Improves Comprehension. *Psychological Science*, *29*(6), 961–971. https://doi.org/10.1177/0956797618754486

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behavior Research Methods*, *48*(4), 1393–1409. https://doi.org/10.3758/s13428-015-0647-3

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants Speech-Language Environments of Children with Cochlear

Implants View project. *Article in Journal of Experimental Psychology Human Perception & Performance*, *26*(5), 1570. https://doi.org/10.1037/0096-1523.26.5.1570

Kadota, S. (2019). *Shadowing as a practice in second language acquisition : connecting inputs and outputs* (S. Kadota (ed.)). Routledge.

Kapnoula, E. C., & McMurray, B. (2016). Newly learned word-forms are abstract and integrated immediately after acquisition. *Psychonomic Bulletin and Review*, *23*(2), 491–499. https://doi.org/10.3758/s13423-015-0897-1

Kapnoula, E. C., Packard, S., Gupta, P., & McMurray, B. (2015). Immediate lexical integration of novel word forms. *Cognition*, *134*, 85–99. https://doi.org/10.1016/j.cognition.2014.09.007

Kapnoula, E. C., & Samuel, A. G. (under review). *Wait long and prosper! Delaying production during training boosts word learning*.

Kapnoula, E. C., & Samuel, A. G. (2019). Voices in the mental lexicon: Words carry indexical information that can affect access to their meaning. *Journal of Memory and Language*, *107*, 111–127. https://doi.org/10.1016/J.JML.2019.05.001

Karpicke, J., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, *319*(5865), 966–968. https://doi.org/10.1126/science.1152408

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

Krashen, S. D. (1985). *The Input Hypothesis: Issues and Implications.* Longman.

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

Lightbown, P. M., & Spada, N. (2013). *How Languages are Learned*. Oxford University Press.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review*, *1*(4), 476–490. https://doi.org/10.3758/BF03210951

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

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. https://doi.org/10.1037//0033-295X.101.4.676

MacDonald, P., & MacLeod, C. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica*, *98*(2–3), 291–310.

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

Magnuson, J. S. (2019). Fixations in the visual world paradigm: where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139. https://doi.org/10.1007/s41809-019-00035-3

Mattys, S. L., & Baddeley, A. (2019). Working memory and second language accent acquisition. *Applied Cognitive Psychology*, *33*(6), 1113–1123. https://doi.org/10.1002/acp.3554

McMurray, B., Clayards, M. A., Tanenhaus, M. K., & Aslin, R. N. (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic Bulletin & Review*, *15*(6), 1064–1071. https://doi.org/10.3758/PBR.15.6.1064

McMurray, B., Kapnoula, E. C. E. C., & Gaskell, M. G. (2016). Learning and integration of new word-forms: Consolidation, pruning and the emergence of automaticity. In M. G. Gaskell & J. Mirković (Eds.), *Speech Perception and Spoken Word Recognition.* (Psychology). https://doi.org/10.4324/9781315772110

McMurray, B., Samelson, V. M., Lee, S. H., & Tomblin, J. B. (2010). Individual differences in online spoken word recognition : Implications for SLI. *Cognitive Psychology*, *60*(1), 1–39. https://doi.org/10.1016/j.cogpsych.2009.06.003

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition*, *86*(2), B33–B42. https://doi.org/10.1016/S0010-0277(02)00157-9

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the Influence of Thematic Fit (and Other Constraints) in On-line Sentence Comprehension. *Journal of Memory and Language*, *38*(3), 283–312. https://doi.org/10.1006/jmla.1997.2543

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory and Language*, *59*(4), 475–494. https://doi.org/10.1016/J.JML.2007.11.006

Mirman, D., Magnuson, J. S., Estes, K. G., & Dixon, J. A. (2008). The link between statistical segmentation and word learning in adults. *Cognition*, *108*(1), 271–280.

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Nora, A., Renvall, H., Kim, J.-Y., Service, E., & Salmelin, R. (2015). Distinct Effects of Memory Retrieval and Articulatory Preparation when Learning and Accessing New Word Forms. *PLOS ONE*, *10*(5), e0126652. https://doi.org/10.1371/journal.pone.0126652

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Nuttall, H. E., Kennedy-Higgins, D., Hogan, J., Devlin, J. T., & Adank, P. (2016). The effect of speech distortion on the excitability of articulatory motor cortex. *NeuroImage*, *128*, 218–226. https://doi.org/10.1016/j.neuroimage.2015.12.038

Ozubko, J. D., & Macleod, C. M. (2010). The Production Effect in Memory: Evidence That Distinctiveness Underlies the Benefit The Production Effect View project. *Article in Journal of Experimental Psychology Learning Memory and Cognition*. https://doi.org/10.1037/a0020604

Ozubko, J. D., Major, J., & Macleod, C. M. (2014). Remembered study mode: Support for the distinctiveness account of the production effect. *Memory*, *22*(5), 509–524. https://doi.org/10.1080/09658211.2013.800554

Page, M., & Norris, D. (1998). The primacy model: a new model of immediate serial recall.

Psychological Review, *105*(4), 761–781.

Page, M., & Norris, D. (2008). Is there a common mechanism underlying word-form learning and the Hebb repetition effect? Experimental data and a modelling framework. In A. Thorn & M. Page (Eds.), *Interactions Between Short-Term and Long-Term Memory in the Verbal Domain* (pp. 148–168). Psychology Press.

Page, M., & Norris, D. (2009). *A model linking immediate serial recall, the Hebb repetition effect and the learning of phonological word forms*. *364*(1536), 3737–3753. https://doi.org/10.1098/rstb.2009.0173

Payton, K. L., Uchanski, R. M., & Braida, L. D. (1994). Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *Journal of the Acoustical Society of America*, *95*(3), 1581–1592. https://doi.org/10.1121/1.408545

Protopapas, A. (2007). CheckVocal: a program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods, 39*, 859–862. https://doi.org/10.3758/BF03192979

Richtsmeier, P. T., Gerken, L. A., Goffman, L., & Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition*, *111*(3), 372–377. https://doi.org/10.1016/j.cognition.2009.02.009

Riley, K. G., & McGregor, K. K. (2012). Noise hampers children's expressive word learning. *Language, Speech, and Hearing Services in Schools*, *43*(3), 325–337. https://doi.org/10.1044/0161-1461(2012/11-0053)

Rodriguez-Fornells, A., Cunillera, T., Mestres-Missé, A., & De Diego-Balaguer, R. (2009).

Neurophysiological mechanisms involved in language learning in adults. In *Philosophical Transactions of the Royal Society B: Biological Sciences* (Vol. 364, Issue 1536, pp. 3711–3735). Royal Society. https://doi.org/10.1098/rstb.2009.0130

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, *17*(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635. https://doi.org/10.1111/j.1532-7078.2010.00033.x

Salverda, A. P., & Tanenhaus, M. K. (2017). The visual world paradigm. In A. M. B. de Groot & P. Hagoort (Eds.), *Research Methods in Psycholinguistics and the Neurobiology of Language*. Wiley.

Samuel, A. G. (1996). Does lexical information influence the perceptual restoration of phonemes? *Journal of Experimental Psychology: General*, *125*(1), 28.

Samuel, A. G., & Larraza, S. (2015). Does listening to non-native speech impair speech perception? *Journal of Memory and Language*, *81*, 51–71.

Scheepers, C., Keller, F., & Lapata, M. (2008). Evidence for Serial Coercion: A Time Course Analysis Using the Visual-World Paradigm. *Cognitive Psychology*, *56*(1), 1–29.

https://doi.org/10.1016/j.cogpsych.2006.10.001

Seedorff, M., Oleson, J., Cavanaugh, J., & McMurray, B. (2017). Eyetracking Analysis in R. *R Package Version 0.1.15*.

Seedorff, M., Oleson, J., & McMurray, B. (2018). Detecting when timeseries differ: Using the Bootstrapped Differences of Timeseries (BDOTS) to analyze Visual World Paradigm data (and more). *Journal of Memory and Language*, *102*, 55–67. https://doi.org/10.1016/J.JML.2018.05.004

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*(2), 833–870.

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. In *Linguistics and Language Compass* (Vol. 3, Issue 1, pp. 236–264). Blackwell Publishing Inc. https://doi.org/10.1111/j.1749-818X.2008.00112.x

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Tanenhaus, M. K., & Trueswell, J. (1995). *Sentence comprehension.*

Thorn, A., & Page, M. (2008). *Interactions Between Short-Term and Long-Term Memory in the Verbal Domain*. Psychology Press.

Trueswell, J. C. (1996). The Role of Lexical Frequency in Syntactic Ambiguity Resolution. *Journal of Memory and Language*, *35*(4), 566–585. https://doi.org/10.1006/jmla.1996.0030

Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science, 167*(3917), 392–393.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 1–14. https://doi.org/10.1037/0278-7393.32.1.1

Yu, C., & Smith, L. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*.

Zamuner, T. S., Morin-Lessard, E., & Strahm, S. (2016). Spoken word recognition of novel words, either produced or only heard during learning. *Journal of Memory and Language, 89*, 55–67.

Zamuner, T. S., Strahm, S., Morin-Lessard, E., & Page, M. (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

Appendix

Table A1. *Means and standard deviations for accuracy and reaction times (RT) at test by Condition, Phase, and Experiment*

Accuracy						
	Phase 1			Phase 2		
	N	*M*	*SD*	*N*	*M*	*SD*
Experiment 1						
Perception-Only	40	97.2%	6.2%	40	99.1%	3.3%
Production	40	97.0%	9.3%	40	99.2%	2.5%
Experiment 2						
Perception-Only	41	94.2%	10.6%	41	99.4%	1.9%
Production	41	93.1%	10.3%	41	99.1%	4.1%
Experiment 3						
Perception-Only	41	94.2%	10.1%	41	99.5%	1.6%
Production	41	95.9%	8.1%	41	99.8%	1.0%

Reaction times (in ms)						
	Phase 1			Phase 2		
	N	*M*	*SD*	*N*	*M*	*SD*
Experiment 1						
Perception-Only	40	1,328	288	40	1,097	290
Production	40	1,328	311	40	1,088	289

Experiment 2						
Perception-Only	41	1,419	336	41	1,154	206
Production	41	1,560	602	41	1,189	220
Experiment 3						
Perception-Only	41	1,416	328	41	1,122	186
Production	41	1,375	301	41	1,127	250

Table A2. *Full 3×2×2×2 ANOVA results*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Experiment	7.263	2	3.631	4.376	.015	0.071
Condition	0.002	1	0.002	0.013	.909	<.01
Phase	16.095	1	16.095	45.665	<.001	0.286
TimeWindow	479.539	1	479.539	1044.344	<.001	0.902
Experiment × Condition	0.045	2	0.022	0.152	.859	0.003
Experiment × Phase	0.488	2	0.244	0.692	.503	0.012
Experiment × TimeWindow	5.583	2	2.791	6.079	.003	0.096
Condition × Phase	0.368	1	0.368	3.107	.081	0.027
Condition × TimeWindow	0.048	1	0.048	0.497	.482	0.004
Phase × TimeWindow	7.319	1	7.319	46.086	<.001	0.288
Experiment × Condition × Phase	0.036	2	0.018	0.150	.861	0.003
Experiment × Condition × TimeWindow	0.083	2	0.042	0.433	.650	0.008
Condition × Phase × TimeWindow	1.091	1	1.091	13.481	<.001	0.106
Experiment × Condition × Phase × TimeWindow	0.017	2	0.009	0.107	.899	0.002
Error		114				

Table A3. *Full results of post-hoc comparisons (Bonferroni-corrected)*

Phase	TW	Mean Difference (Perception-Only – Production)	*SE*	*p*	Lower Bound	Upper Bound
1	Early	.017	.024	.465	-.030	.064
1	Late	-.090	.060	.131	-.209	.028
2	Early	-.040	.021	.059	-.082	.002
2	Late	.125	.054	.022	.019	.232

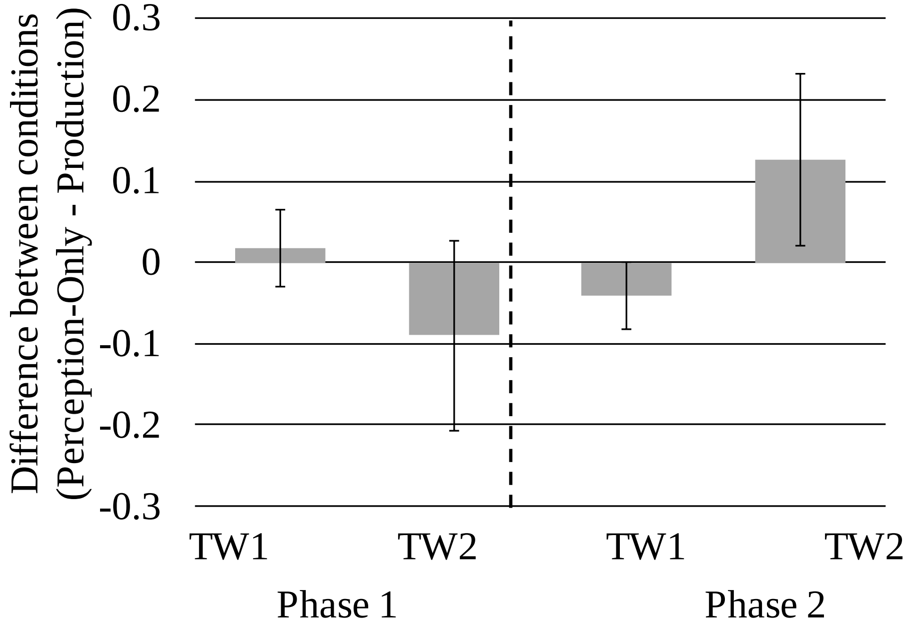

Figure A1. Average differences of looks to the target (empirical-logit-transformed) per testing phase (Phase 1/Phase 2), and time window (TW) within trial (early: TW1/late: TW2) across experiments. Error bars show 95% confidence intervals of the mean differences. *Note*. This alternative visualization of the ANOVA results presents empirical-logit-transformed data, as used in the statistical analyses.