Wait long and prosper!

Delaying production alleviates its detrimental effect on word learning

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: DELAYING PRODUCTION HELPS WORD LEARNING

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

Recent work by Baese-Berk and Samuel (2022) suggests that immediate –but not delayed– production has a detrimental effect on learning a non-native speech sound contrast. We tested whether this pattern is also found for word learning. Each participant learned 12 new words in one of four training conditions: *Perception-Only*, *Immediate-Production*, *2-seconds-Delayed-Production*, and *4-seconds-Delayed-Production*. At test, we assessed how well new words were embedded into the mental lexicon by measuring the degree to which they could drive phonemic recalibration (also called "perceptual learning"). Training and testing were repeated on the next day along with a word recognition task assessing lexical configuration. Replicating previous findings, Day 1 results showed that repeating a new word immediately after hearing it disrupted learning compared to just hearing it. Critically, in line with our prediction, this negative effect disappeared when a 4-second pause was inserted between hearing and producing each word.

Keywords: word learning, production, spoken word recognition, mental lexicon

Introduction

Imagine that you are at a conference, and a colleague introduces you to several people. You, of course, want to make sure you will later remember each person's name. A common strategy would be to say each person's name, perhaps when your colleague first mentions it. In fact, a well-known technique for remembering a name when you meet a person is to respond "Nice to meet you XXXX"; saying "XXXX" aloud should help you remember the person's name.

This mnemonic technique is quite intuitive, reflecting the common-sense belief that producing a name/word will help us to encode it in memory. This intuition is supported by a substantial literature, with the pattern being robust enough to have its own name – the "Production Effect". The original work in this domain goes back at least to papers by Conway and Gathercole (1987; Gathercole & Conway, 1988), with a substantial body of work being done by MacLeod and his colleagues (e.g., MacLeod et al., 2010) and more recently by Mama and Icht (e.g., 2016, 2018). In the vast majority of papers on the Production Effect, participants are given a set of printed words to memorize, followed by a final recall or recognition task. The basic result is that items that are said aloud during learning are better remembered than those that are just read (with the effect primarily found on within-subject tests; e.g., MacLeod, 2011; MacLeod et al., 2010).

Although the Production Effect is both intuitive and empirically well-established, there is now a growing body of evidence showing that, rather than helping, speech production can actually be ***detrimental*** to learning under certain circumstances. This counter-intuitive result has been found when people are either learning new words (e.g., Leach & Samuel, 2007), or are learning a phonetic contrast that is not present in their native language (L1; e.g., Baese-Berk & Samuel, 2016, 2022). These situations contrast with the task in a typical Production Effect

experiment, as in the latter case the items are words that the participants already know and the task is to activate existing lexical representations for later recall or recognition. In contrast, in the present work the focus is not the effect of production on the memorization of familiar items, but rather on its role in establishing *new* representations.

Even when the task is to learn new words or to learn a new phonetic contrast, a detriment in learning due to production only occurs under certain circumstances. These circumstances can provide a window into the underlying processes that support learning new words or new phonetic contrasts. The primary goal of the current study is to gain a better understanding of these underlying processes by showing how the production-driven deficit in word learning depends on the *timing* of the production requirement.

To situate our experiments properly, we start by providing a short summary of the growing literature in which production-driven deficits have been observed. We will first describe the key findings for learning of new words, and then describe the relevant results when new phonetic contrasts are being learned. It is worth noting that repeatedly observing production-driven deficits at two very different levels of language acquisition (i.e., word-level and phoneme-level) suggests that the effect stems from some fundamental properties of the language system, rather than being just a quirk that affects a narrow piece of the system.

Detrimental effects of production in learning words and speech sounds

The earliest finding of a production cost during word learning that we are aware of was reported by Leach and Samuel (2007). We describe this study in some detail because the procedures are very similar to those in the current study. The authors used several different training regimes and several different measures of word learning. We focus here on their

Experiments 2 and 5, those using procedures closest to the current study. American English listeners were taught 12 new words by associating each word with the picture of an unfamiliar object. The words were all either three-syllables long (e.g., "gatersy") or four-syllables (e.g., "penivasher"). In Experiment 2, in each training trial the participant heard one of the words and simultaneously saw two of the unusual objects, side by side. The participant used the left or right button on a keypad to indicate whether the word matched the picture on the left or on the right. Immediate feedback was provided by keeping only the correct picture on the screen for one second after the response. In Experiment 5, everything was identical except that the participant was also required to produce the word that was said after making the left/right picture choice.

Leach and Samuel (2007) tested word learning with multiple measures and used the results to outline a distinction between "lexical configuration" and "lexical engagement". Lexical configuration is assembling the information that comprises a lexical entry – what a word sounds like, how it is spelled, its meaning, its syntactic role, etc. Lexical engagement refers to processing-related aspects of a lexical entry. For example, in many models, an active lexical entry will affect the activation of other representations: words with similar meanings may be activated (via spreading activation), words that sound similar may be inhibited (via lateral inhibition), and sublexical units that are consistent with the active lexical representation may be facilitated (via top-down feedback). Broadly speaking, initial learning of a word may be more about building its lexical configuration, with lexical engagement only developing when a lexical representation is more fully established. In some cases, the development of the two aspects may overlap, as suggested by work showing immediate lexical engagement of newly learned words (Kapnoula et al., 2015; Kapnoula & McMurray, 2016).

One of the post-learning measures that Leach and Samuel (2007) used was primarily aimed at lexical configuration, and a second was primarily intended to tap lexical engagement. To measure configuration, the authors presented words and pseudowords in decreasing levels of white noise, and had subjects report an item when the noise level was low enough to permit recognition. On this task, which involved recognizing the sequence of segments, production during training had a positive effect; participants who had produced words during the learning phase could recognize those words at higher noise levels than those who had only listened to them.

To measure engagement, Leach and Samuel tested how well the newly learned words could support *phonemic recalibration* (Norris et al., 2003). In recalibration studies, a number of words are systematically mispronounced, and these systematic mispronunciations cause a listener to retune a phonemic boundary. For example, Norris et al. initially presented one set of listeners with words in which each occurrence of /s/ was realized as an ambiguous mixture of /s/ and /f/; for a different set of listeners, each /f/ was realized with the ambiguous /s-f/ mixture. After exposure, listeners in the /s/ group expanded their /s/ category (hearing ambiguous stimuli as more /s/-like), while listeners in the /f/ group did the reverse. A widely accepted interpretation of this finding is that the lexical context biasing /s/ or /f/ during the exposure phase engages with the sublexical representations for /s/ or /f/, supporting their retuning.

Because recalibration only occurs when the ambiguous sounds occur in real words (Norris et al., 2003 showed that pseudoword contexts do not work), the recalibration effect can be used to assess lexicality. Indeed, Leach and Samuel used this paradigm to test whether the newly learned words had acquired the property of lexical engagement. The recalibration test was similar to that of typical recalibration experiments. The critical difference was that participants first learned a

set of new words, with or without producing them, and then, during the exposure phase, the ambiguous sounds were embedded in those novel words. The results indicated that participants who had *not* produced the words during learning showed the recalibration effect, but those who had produced the words during learning did not. This was the first reported case we are aware of in which producing a word during the learning phase blocked an important part of lexical acquisition.

A study by Kaushanskaya and Yoo (2011) provides an initial clue about the conditions that lead to a production-driven deficit in learning new words. The authors had listeners learn to associate novel monosyllables or bisyllables with English translations. For example, a listener might learn that "weg" meant "table". In one experiment, all of the novel words were sequences that were built by drawing from a set of four English vowels and four English consonants. One group just listened to the new words while learning the associations, whereas a second group was required to produce the word on each learning trial. Under these conditions, production helped – learning was better for the group that produced the words. In a second experiment, all of the procedures were the same, but the novel words were built from a set of four vowels and four consonants in which half of the phonemes were English, and half of the vowels and half of the consonants were non-native. Under these conditions, the results flipped, with the Production group showing a learning deficit compared to the participants who only listened on each learning trial. This result suggests that one potentially important factor in observing a production-driven deficit is the difficulty of the material to be learned.

A final set of studies used eye-tracking measures to examine the impact of production during the learning of new words. Zamuner et al. (2016) provided brief training for young adults to associate eight new words with drawings of nonce-animals. Half of the words were just heard

(twice) during each training trial, while the other half were heard once and then produced. In a following test phase, two of the pictures were shown on a screen when one of the new words was presented auditorily. Eye-tracking measures indicated that words that had been produced during training were better able to direct looks to the correct visual stimulus.

This example of production facilitating the learning of new words can be contrasted with two studies that were modeled on Zamuner et al. (2016), but that instead found production-driven deficits, rather than facilitation. Zamuner et al. (2018) used a slightly simplified version of the original procedure in order to be able to test 4-6 year-old children, rather than young adults. In this case, the eye-tracking measures showed that words learned with production were *less* able to direct eye movements to the learned referents.

Kapnoula and Samuel (2022) conducted a set of eye-tracking experiments that were modeled on the Zamuner et al. (2016) study, with adult participants. The key difference was in the amount of training that was provided to learn the new words. There was an initial training phase that, like Zamuner et al., only included two training trials for each new word with its visual associate. A test using eye-tracking, after this initial training, replicated the positive effect of production during training. However, training then continued, with each word-picture association presented an additional ten times. When the eye-tracking test was conducted after this more thorough learning, words that had been trained with production were *less* able to direct eye movements to the corresponding visual referents. These results suggest that at the very earliest moments of learning a new word, having some additional cue (in this case, production) is useful, but after this very early stage production is detrimental to learning the new word. The early advantage might occur because newly acquired representations are so weak at first that any additional (e.g., articulatory) information may enrich them, bootstrapping early learning (e.g., Mattys &

Baddeley, 2019). However, once lexical representations have been minimally stabilized, production seems to hurt further integration into the mental lexicon.

Overall, there is a growing body of research, using a range of training conditions and dependent measures that demonstrates a negative impact of producing words during the learning phase. As noted above, this counter-intuitive pattern is not limited to the case of learning new words; it has also now been found in a series of studies that have measured the acquisition of non-native phonemic contrasts. As with word-level learning, several stimulus types and training regimes have been used. For example, Baese-Berk (2019) had American English listeners learn a distinction between prevoiced and voiced stop consonants. This is a phonemic contrast in various languages, including Hindi, but it is not present in American English – prevoiced tokens are not distinguished from those with short positive voice onset times (VOT) (i.e., /b/, /d/, or /g/ can all be realized either way). Baese-Berk exposed listeners to tokens that varied in VOT in two different ways: Half of the listeners received a bimodal distribution, with many substantially prevoiced tokens and many tokens with VOTs closer to zero, with few in between. For the other half of the listeners, the distribution was closer to Gaussian, with relatively few tokens from either the low end or the high end. When participants just listened to the tokens, those exposed to the bimodal distribution developed two distinguishable categories, the prevoiced versus short lag sounds, reflected in improved performance on between-category pairs on a following discrimination test (see also Maye et al., 2002). In contrast, when participants were required to produce each token as it was heard, there was no learning of the two separate categories, even for those receiving the bimodal distribution – production blocked the learning of the perceptual distinction.

Baese-Berk and Samuel (2016, 2022) have demonstrated a comparable result for native Spanish listeners who were trained on a distinction that is used in languages such as English and Basque but not in Spanish – /s/ versus /ʃ/ (Spanish has /s/, but not /ʃ/). In multiple experiments, Spanish listeners were given explicit discrimination training on the distinction, with feedback. The recurring result in these experiments has been that production interferes with learning to perceive the phonemic contrast – people who just made a perceptual judgment on each training trial learned the distinction, but those who also produced the sounds did not. The set of experiments in this domain includes various manipulations to try to specify the conditions and mechanisms that lead to a detrimental effect of production. In the current study we pursue these questions at the word level. We begin by describing different possible mechanisms that could underlie these detrimental effects.

Why does production hurt learning?

A core question, for both learning new words and learning a new phonemic distinction, is *why* production can impair perceptual learning. To date, three possibilities have been considered; these three are not mutually exclusive (see also Wright, 2021 for a discussion of possible mechanisms at both the cognitive and neural levels). Perhaps the most intuitive possibility is that when participants produce unfamiliar words or unfamiliar sounds they mispronounce them, and in so doing, give themselves perceptual input that is not correct. Such incorrect input could plausibly disrupt learning of the correct perceptual stimulus. Although this could be one factor, there are at least two reasons to doubt that it plays a major role. At the level of learning new words, the stimuli and conditions make it very unlikely that participants will produce many such mispronunciations. For example, in the Leach and Samuel (2007) study, in which people learned items like “gatersy” or “penivasher”, it is conceivable that they might have made a small error on

the first one or two learning trials for an item, but each item was presented 120 times during training. As such, the self-productions were almost certainly all accurate for virtually all of the trials. At the level of learning a new phonemic contrast, Baese-Berk (2019) coded all of the productions, and found that accuracy of production did not predict the level of perceptual learning. Rather, the variability of production was related to the level of learning that was achieved: People whose productions were relatively stable in terms of VOT tended to achieve better learning, whether or not their VOT was a good match to the input.

Thus, while we cannot rule out any role for bad self-input as the source of a production disadvantage, it is unlikely to be a major factor. The two other possibilities that have been considered are (a) learning a perceptual distinction may rely on representations that compete with representations that support learning a new production, and (b) engaging in production, at the moment when perceptual learning would be taking place, leads to a lost opportunity to learn. For learning the non-native /s/-/ʃ/ distinction, Baese-Berk and Samuel's (2016, 2022) research program has tested both of these possibilities, and found support for both.

To test whether there might be some incompatibility between the to-be-learned perceptual representations and the to-be-learned production representations, they compared phonemic category learning under two different production requirements. In one case, participants had to produce the syllable they heard on a given training trial (i.e., the token of "sa" or "sha"), while in the other case they instead had to produce the name of a letter that was displayed on a screen at that moment. Both production conditions led to worse perceptual learning than the perception-only condition, but producing the name of an unrelated letter was less disruptive. This difference suggests that the additional disruption reflects an incompatibility between the perceptual and production codes being learned for the /s/-/ʃ/ contrast.

The "lost opportunity" account comes from viewing a perceptual learning regime that includes production as essentially a dual-task situation: The participant is required to both learn the perceptual distinction and produce the required sound. Particularly when the contrast involves an unfamiliar sound (such as a prevoiced stop for American English speakers, or /ʃ/ for Spanish speakers), the production task may require some attentional focus (or other form of cognitive resources). To the extent that the participant must focus on production, there is less attention available to learn the perceptual distinction (or new word). To test this possibility, Baese-Berk and Samuel (2022) compared the standard production case, in which the production is done just before or just after the perceptual judgment, to conditions in which the production requirement was delayed. Consistent with the view that the production-driven deficit reflects the withdrawal of attention from perceptual encoding, delaying the production requirement by two seconds reduced the deficit, and delaying it by four seconds completely abolished the deficit in perceptual learning (see Figure 1).

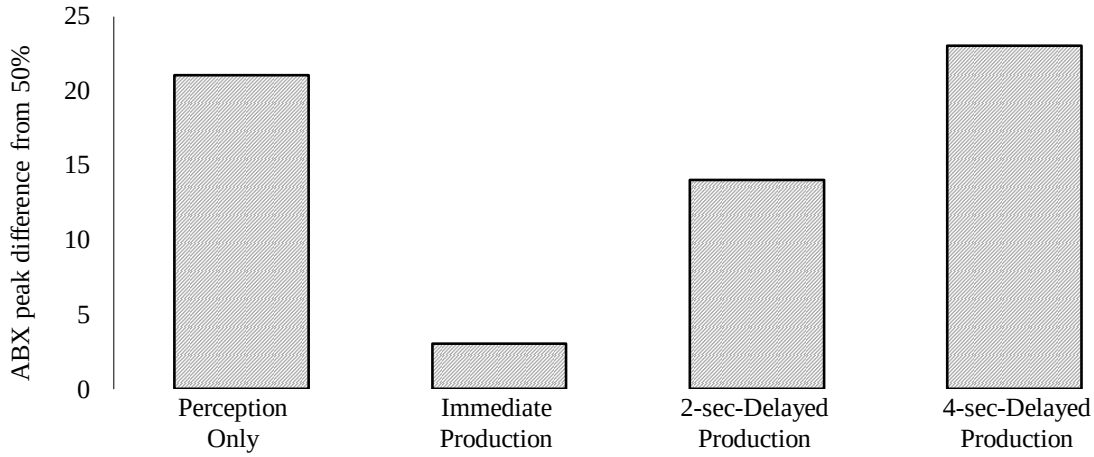


Figure 1. ABX peak difference from chance (50%). Note: In this paradigm, a higher peak reflects better learning of the new phonemic contrast. Based on data reported in Baese-Berk and Samuel (2022).

Present study

In the current study, we test whether the temporal manipulation of the production requirement has a similar effect in the context of learning new words. That is, if the production cost reflects a fundamental aspect of acquiring new language units, regardless of level, then we should expect that there will be comparable effects of delaying production when the learning is at the lexical level. Alternatively, it could be the case that what is needed during the learning of a new word is different than what is needed when learning a new phonemic contrast. Finding comparable effects of delaying production would suggest that the basic properties of establishing representations are the same at both the lexical and sublexical levels; different effects would point to different underlying learning routines.

To address this question, we used an experimental design similar to that of Leach and Samuel (2007), in which lexical integration is assessed in terms of the ability of new words to drive phonemic recalibration. As in the original study, we used a *Perception-Only* and an *Immediate-Production* training condition. Comparing these two conditions provides a conceptual replication of the baseline detrimental effect of production found by Leach and Samuel (2007). In addition, we included two new training conditions, in which participants were asked to produce each item, but critically, the production itself was delayed. Specifically, we tested two delays: two (2) seconds and four (4) seconds (corresponding to the *2-seconds-Delayed-Production*, and *4-seconds-Delayed-Production* condition, respectively). The decision to use these delays was based on previous work showing that a useful representation of acoustic information can be

maintained for about this much time[1] (Darwin et al., 1972). These delays also match the time course tested for new phonemic contrasts in Baese-Berk and Samuel (2022).

In addition to this primary goal, we also addressed two secondary questions. The first was whether a potential effect of production would be modulated by off-line consolidation. Previous work has provided robust evidence that sleep-driven consolidation plays a critical role in lexical integration (Davis & Gaskell, 2009; Dumay & Gaskell, 2007, 2012; Tamminen et al., 2010). For example, in Leach and Samuel (2007) training and testing were repeated over a period of five days and it was found that the strength of the lexical representations gradually increased (see also Gaskell & Dumay; 2003, for a similar multi-day training manipulation). Thus it was possible that any effects related to lexical integration would not appear on the first day. In anticipation of this possibility we repeated the procedure (training and testing) on the next day. Here we note that if we were interested in the effect of consolidation per se, we would not include more training at the beginning of the second session. Our primary goal was rather to maximize the chance of getting a recalibration effect at all.

Our other secondary issue involved lexical configuration. Even though we were primarily interested in the effect of delayed production on lexical engagement, we also collected an independent measure of lexical configuration (the *recognition task* described below). The decision to collect a separate lexical configuration measure was based on the findings of Leach and Samuel (2007), who found that even though production hurt lexical engagement, it seemed to have a facilitatory effect on lexical configuration. To avoid contamination of our primary

[1] Recent work has cast doubt on whether acoustic information can be maintained beyond two seconds (Caplan et al., 2021). However, this issue is still very much a matter of debate. For example, Sarrett et al. (2020) showed that listeners retain subphonemic information for approximately ~900 ms even when they do not expect further disambiguating information. Regardless of how this debate is resolved, the delays that we used here produced interpretable and interesting results.

measure of lexical engagement, we had participants perform the recognition task at the very end of the second session (i.e., second day).

We conducted an initial experiment (Experiment 1) to ensure that our test stimuli would allow us to measure phonemic recalibration driven by familiar words (i.e., to replicate the basic recalibration effect; Norris et al., 2003; Samuel & Kraljic, 2009). Establishing the baseline recalibration effect was a critical step, as the effect depends on a delicate stimulus manipulation and rigorous piloting. Experiment 1 allowed us to verify that our procedures and test stimuli were suitable for assessing phonemic recalibration. With this established, we conducted Experiment 2, in which we measured phonemic recalibration driven by newly learned words. Crucially, in Experiment 2, each participant was assigned to one of four training conditions: Perception-Only, Immediate-Production, 2-seconds-Delayed-Production, and 4-seconds-Delayed-Production.

Experiment 1

Method

Participants

Thirty-eight (28 females; mean age = 27.5 years) native speakers of Spanish participated in Experiment 1. Most participants were also fluent in Basque, which was foreseen and taken into account in selecting the stimuli (see ***Materials*** below). All participants self-reported having 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 and procedure

Participants were given two tasks. The first task served to expose listeners to critical words that either included a target /f/ sound (e.g., /f/ in *biografía* [=biography]), or a target /s/ sound (e.g., /s/ in *contraseña* [=password]). During the exposure task the target sounds were replaced with an ambiguous /f-s/ sound. After this exposure phase, the participants' second task was a 2AFC (f/s) categorization test that was used to measure phonemic recalibration.

Exposure to mispronunciations: Old/new task. The purpose of this task was to expose participants to known words containing ambiguous /f-s/ sounds; half of the participants were exposed to words containing a mispronounced /f/ sound and the other half were exposed to words containing a mispronounced /s/ sound. To expose participants to ambiguous f/s utterances, we used an old/new recognition task similar to that of Leach & Samuel (2007). This task avoids any difficulty listeners may have in categorizing ambiguous stimuli as word versus nonword (i.e., lexical decision, the most commonly used exposure task).

Immediately before the exposure phase, participants heard a list of 18 items, nine words and nine nonwords, and were instructed to memorize them, as they would be later tested on them. The block of 18 items was presented three times. Stimuli were presented auditorily in a random order and participants were asked to press the spacebar to advance from one item to the next. Participants had 5,000 ms to respond and the next trial started 600 ms later.

Next, in the exposure phase, participants were presented with a second list of auditory stimuli consisting of: (a) the 18 items of the first list ("old" trials), (b) 12 critical new words – six containing an /f/ sound (e.g., *biografía*) and six containing an /s/ sound (e.g., *contraseña*; see Table 1), and (c) 24 word and 24 nonword fillers (filler "new" trials). None of the 18 "old"

words or the 48 filler items contained an /f/ or an /s/ (see Appendix Table A1). Critically, for half of the participants, the six /f/ sounds were replaced by ambiguous /f-s/ sounds, and for the other half, the six /s/ sounds were replaced by ambiguous /f-s/ sounds.

Table 1. *List of 12 real words with an /s/ or /f/ sound*

Words with /f/	Words with /s/
biografia [biography]	parisino [parisian]
califato [caliphate]	comisario [commissar]
infable [ineffable]	marquesado [marquise]
albufera [coastal lagoon]	queroseno [kerosene]
clorofila [chlorophyll]	golosina [candy]
periferia [periphery]	contraseña [password]

Note: English translations in parentheses

While listening to each of the items in the second list, participants saw two words on the screen; "nueva" [=new] on the left, and "vieja" [=old] on the right, and their task was to press the corresponding key on a button box to indicate if an item had been played in the first list or not (i.e., whether it was "old" or "new"). Participants had 5,000 ms to respond and the next trial started 600 ms later. The participants' task was irrelevant to our experimental manipulation; the goal was simply to expose them to the critical items.

Each item from the first list occurred three times in the second list, resulting in a total of 54 "old" trials. Each of the critical 12 items also occurred three times, resulting in 18 correctly pronounced and 18 mispronounced trials. Lastly, each filler was presented twice, for a total of 96 filler trials. Thus, there were 132 "new" trials (36 experimental + 96 fillers). The 186 tokens were presented in a random order. For our purposes, the 18 mispronounced critical tokens (6

items, each presented 3 times) were what mattered; the other trials were simply a way to embed those trials in a larger context.

Phonemic recalibration test: ufi/usi task. After the exposure phase, participants did a 2-alternative-forced-choice (2AFC) task. On each trial, they heard an auditory stimulus from a seven-step /ufi-usi/ continuum and saw two nonwords on the screen; "ufi" on the left, and "usi" on the right. Their task was to categorize each item as /ufi/ or /usi/ by pressing the corresponding key on a button box. Participants had 2,500 ms to respond and the next trial started 1,000 ms later. The seven items were presented in 10 different random orders.

Materials

All audio recording, preprocessing, and editing was done using Praat (Boersma & Weenink, 2016). Stimuli were constructed from recordings of natural speech spoken by a female native Spanish speaker in a sound-attenuated room, sampling at 44,100 Hz. We collected multiple recordings and chose one per item based on sound quality. Chosen recordings were cut, cleaned (background noise and occasional click/pop sounds removed), and intensity-scaled to 70 dB.

Old/new task: All real word and nonword items had four syllables. In addition, all nonwords were checked by a Spanish-Basque bilingual research assistant to make sure that they were nonwords in both Spanish and Basque, but morphologically and phonotactically consistent with Spanish.

To construct the mispronounced stimuli, we recorded a correct pronunciation and a mispronunciation of each critical item (e.g., "*contra**s**eña*" and "*contra**f**eña*") and extracted the /f/ and /s/ sounds. The two sounds were then used as endpoints to make a 15-step /f/-to-/s/ continuum and each step was inserted into each of the two contexts (e.g., "*contra**f**eña*" and

"*contraseña*"). Finally, six native Spanish speakers listened to all 360 items (12 words × 15 steps × 2 contexts) and helped us to select the most ambiguous version of each word.

2AFC usi/ufi task: To create the /f/-/s/ continuum the same talker recorded clear tokens of /ufi/ and /usi/. We extracted the /f/ and /s/ sounds, and created various mixtures of the two. The sounds were mixed in 5% increments, e.g., 95% /f/ mixed with 5% /s/, 90% /f/ with 10% /s/ and so forth. Based on pilot data, we selected seven stimuli: 10%, 20%, 30%, 40%, 50%, 60%, and 70% /s/. Each of these was inserted into the /ufi/ context, replacing the original /f/ sound.

Results

The responses from one participant were not available due to technical difficulties.

We calculated the average proportion of "ufi" versus "usi" responses for each of the seven items of the /ufi-usi/ continuum. Figure 2 shows the average proportion of "usi" responses per step. The solid curve shows the results for participants who were exposed to ambiguous /s/ sounds, and the dashed curve shows the corresponding data for those who had heard ambiguous /f/ sounds during the old/new task. There is a clear phonemic recalibration effect; participants exposed to ambiguous /s/ sounds were more likely to classify ambiguous /u?i/ items as /usi/ than subjects who had heard ambiguous /f/ sounds.

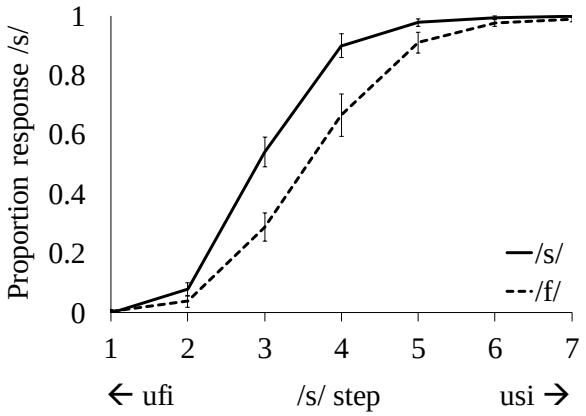


Figure 2. Average proportion of /usi/ responses per /ufi-usi/ step by exposure condition in Experiment 1. Note: Error bars indicate standard error of the mean.

To test this effect statistically, for each participant we computed the average proportion of /usi/ responses across the middle three steps of the continuum (Bertelson et al., 2003; Leach & Samuel, 2007; Samuel, 1986). Raw proportions were empirical-logit-transformed and an independent-samples t-test was used to compare the resulting values between exposure groups. The 19 participants who were exposed to ambiguous /s/ sounds were more likely to classify ambiguous stimuli as /usi/ (M = 81%) than the 18 participants exposed to ambiguous /f/ sounds (M = 62%), $t(35) = 3.78$, $p < .001$.

Discussion

The exposure procedure produced clear evidence for lexically-driven phonemic recalibration. That is, participants who were systematically exposed to mispronounced /f/ sounds learned to perceive ambiguous /f-s/ sounds as /f/, whereas participants exposed to mispronounced /s/ sounds learned to perceive the same ambiguous /f-s/ sounds as /s/. This pattern of results verifies that our procedure can drive phonemic recalibration, and that our test continuum is suitable for measuring this effect.

Experiment 2

Experiment 2 examines how immediate and delayed production of novel words affects the degree to which these words are integrated into the mental lexicon. First, we assessed word learning in the absence of production (i.e., Perception-Only training condition). This condition was treated as a baseline against which each of the three production conditions was compared (i.e., Immediate-Production, 2-seconds-Delayed-Production, and 4-seconds-Delayed-Production). The degree of lexical integration that was achieved in each training condition was assessed based on whether novel words could drive phonemic recalibration.

Method

Participants

One hundred and forty nine (149; 103 females; mean age = 25.4 years) native Spanish speakers participated in Experiment 2. The goal was to have 18 participants per experimental cell (i.e., for each of the eight combinations of Training Condition × Exposure). This number (18) was chosen as an appropriate sample size because it is well within the range of sample sizes typically used in recalibration experiments. Experiment 2 was identical to Experiment 1 in terms of participant characteristics, compensation, and ethical approval procedures.

Design and procedure

As in Experiment 1, Experiment 2 included an ambiguous /f-s/ exposure task followed by a 2AFC /ufi-usi/ task to measure phonemic recalibration. The critical difference between the two experiments was that in the exposure task of Experiment 2 the ambiguous /f-s/ sounds were embedded in *newly learned* words. In the exposure phase, participants did a 5AFC task instead of an old/new task. This was done because presenting newly learned words could cause

confusion as to whether a word counts as new or not. In addition, like the old/new task, the 5AFC exposure task avoids potential categorization problems that listeners might have doing lexical decision on acoustically ambiguous stimuli.

Experiment 2 incorporated two features that were mentioned in the Introduction. First, we included additional training and testing to determine whether the results would be affected by sleep-driven consolidation and additional training. Given the well-documented boosting effects of sleep-driven consolidation in word learning, repeating the training and testing procedure the next day allowed us to maximize the chances of observing lexical integration. For this purpose, we had participants return to the lab the next day to repeat all tasks. Second, at the end of the second day, participants performed an old/new recognition task that was designed to measure lexical configuration of new words. As mentioned in the Introduction, this test was motivated by Leach and Samuel's (2007) finding that production during word learning may facilitate configuration, even under circumstances in which it impairs the development of lexical engagement. In sum, Day 1 data were used to address our main question, while Day 2 data were used to address secondary questions related to consolidation and lexical configuration. Table 2 shows the order of the tasks on each day of the experiment.

Table 2. *List of phases and corresponding tasks of Experiment 2*

Order	Phase	Task
1	Novel word learning	2AFC picture ID
2	Exposure to mispronunciations	5AFC picture ID
3	Phonemic recalibration test	2AFC ufi/usi
4	*Lexical configuration test	Word recognition

* Lexical configuration was only assessed at the end of the second day

Novel word learning: 2AFC picture identification task. In the first task, participants were trained on twelve new words (see Table 3, in the *Materials* section below, for a list); six contained an /f/ sound (e.g., ranpofita) and six an /s/ sound (e.g., nultosera). A picture of an unusual object was randomly assigned to each novel word as its referent and participants were asked to learn these associations in a picture identification task. For this task, each participant was assigned to one of the four training conditions: Perception-only, Immediate-Production, 2-seconds-Delayed-Production, and 4-seconds-Delayed-Production. The details of the task varied according to the training condition.

On each **Perception-Only** training trial, participants saw two pictures on the screen and 250 ms later they heard one of the novel words over headphones (see Fig. 3A). Right after the word offset, a question mark appeared on the screen prompting participants to respond. Their task was to report which of the two objects was the referent of the word they heard by pushing the corresponding button on a response box (i.e., left or right). When they responded, the correct picture remained on the screen for 750 ms (providing feedback for learning the association), while the other one disappeared.

For participants in the **Immediate-Production** condition, a similar procedure was followed (see Fig. 3B). Again, participants were asked to wait for the prompt (i.e., the question mark), but in this case they were asked to repeat the word before pushing the corresponding button. The two **Delayed-Production** conditions were similar to the Immediate-Production condition, but the prompt to respond appeared two or four seconds after the word offset (see Figs. 3C and 3D).

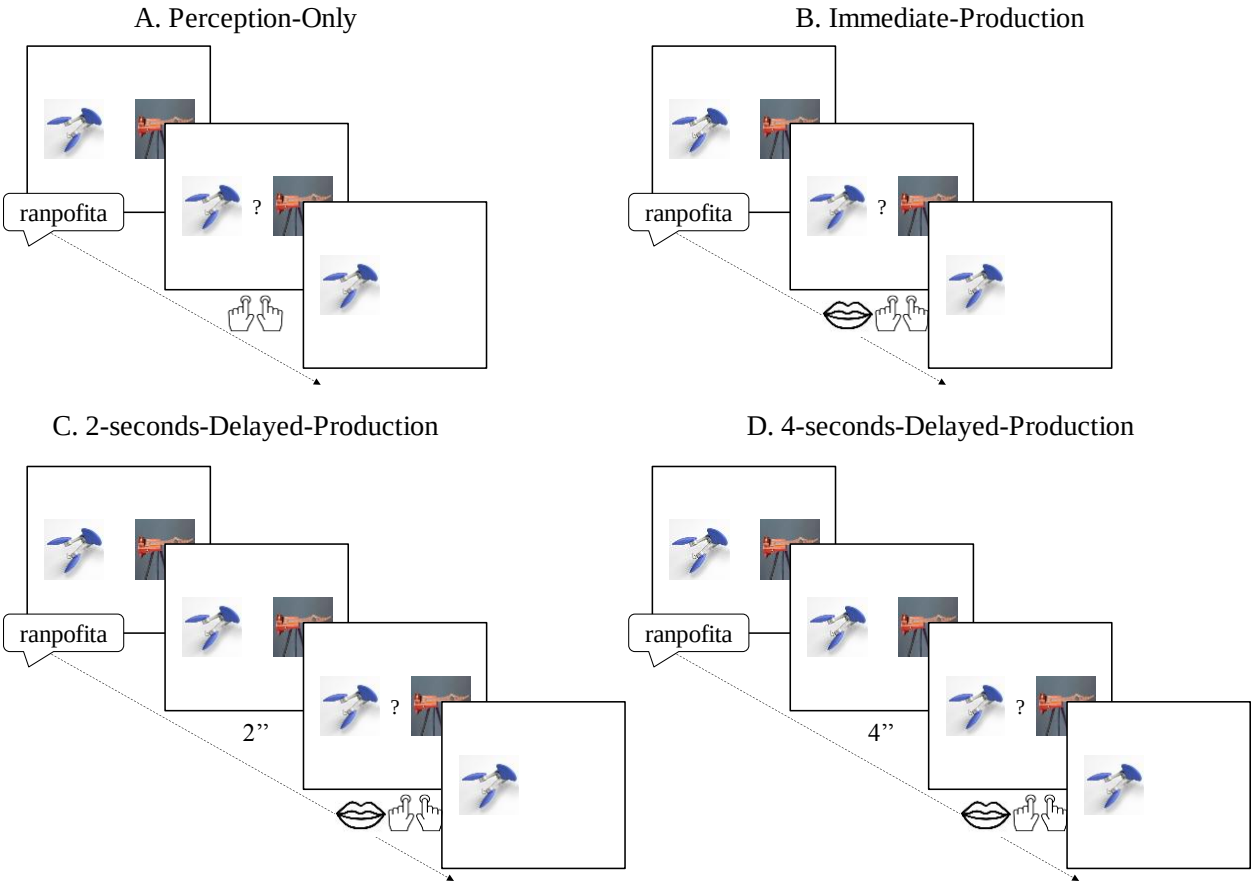


Figure 3. Examples of 2AFC picture identification trials used in the training phase.

Each word was presented 20 times in each training session (i.e., on each day). Each time its referent was randomly paired with one of the other 11 pictures, which served as the lure for that trial. Position of the target versus lure picture (left versus right) was randomly determined on each trial.

Exposure to mispronunciations: 5AFC picture identification task. The purpose of this task was to expose participants to items containing ambiguous /f-s/ sounds. Unlike Experiment 1, the mispronounced items were newly learned words. Half of the participants were exposed to novel words containing a mispronounced /f/ sound; the other half heard novel words containing a

mispronounced /s/ sound (counterbalanced within each of the eight participant subgroups; i.e., 4 training conditions × 2 stimulus groups).

Auditory stimuli were: a) the 12 novel words and b) 24 real word fillers. Half of the 12 novel words were presented with an ambiguous f/s mixture, with half of the participants hearing these ambiguities in /f/ words, and the other half hearing them in /s/ words. The other six novel words were heard with the intact "other" sound. Each of the 36 items was presented as the auditory target in three trials, for a total of (36 items × 3 repetitions) 108 trials. The participants' task was to identify its visual referent on the screen, or report its absence. As in Experiment 1's exposure task, the procedure resulted in each participant hearing 18 mispronunciations – six items, presented three times.

For the 12 novel words, the same visual referents were used as in the training task. For the 24 real words, only half were assigned a visual referent (*visible fillers*). For the other 12 real words, the auditory stimulus never matched any of the pictures on the screen (*invisible fillers*). Each experimental item containing an /f/ sound (e.g., ranpofita) was paired with an experimental item containing an /s/ sound (e.g., nultosera) making a total of six pairs. Then, each of these pairs was grouped with four real words (two visible and two invisible fillers) creating six 6-item sets. The goal of these groupings was to minimize phonological similarity and semantic relatedness among items in any given set. Only items from the same set were ever presented in the same trial and these sets were the same for all participants (see 5AFC sets in Appendix Table A2).

At the beginning of each trial, participants saw four pictures on the screen along with a circle in the center containing the words "NO ESTÁ" [=does not exist] (see Fig. 4). After 250 ms they heard one word over headphones. Participants had to click on one of the four pictures to indicate

which object was the referent of the word they heard. If a participant wanted to indicate that none of the objects corresponded to what they heard, they could click on the circle at the center. We included this option in case a participant heard a mispronounced novel word as being different enough from the learned version to make selecting the referent unacceptable. The "invisible filler" items were included to avoid calling attention to the mispronounced critical items.

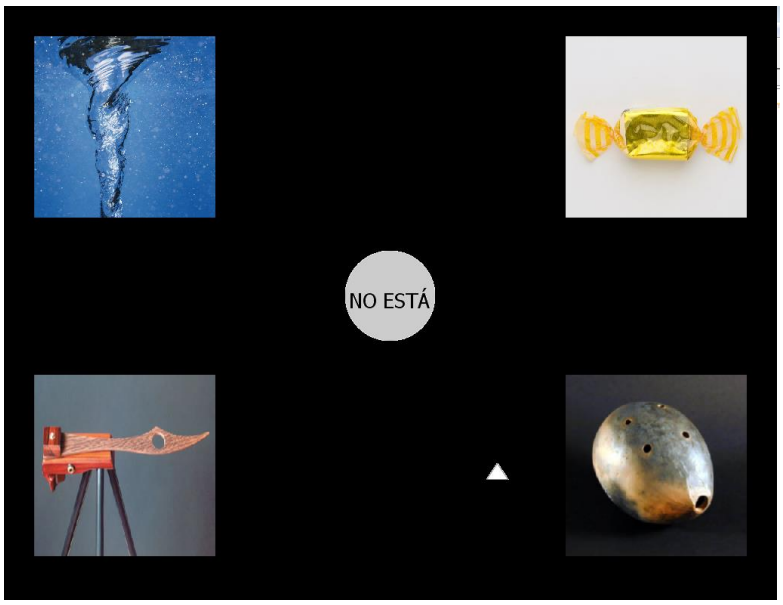


Figure 4. Example of display in the 5AFC task of Experiment 2

Phonemic recalibration test: 2AFC ufi/usi task. This task was identical to that of Experiment 1.

Lexical configuration test: Word recognition task. At the end of the second session (i.e., second day) participants performed a word recognition task. On each trial, they listened to either a newly learned word (e.g., ranpofíta), or a new item that differed from that new word by one consonant – either the one immediately preceding the /f/ or /s/ sound (e.g., rantofita), or the one immediately following it (e.g., ranpofila). Participants listened to each of the 36 items (i.e., 12

targets plus 2 lures per target) and reported whether it was one of the newly learned words, or not. They performed this task twice – first with background noise[2] and then without it.

Materials

All audio recording and pre-processing steps were identical to those of Experiment 1.

2AFC picture identification task. The 12 novel word items were four syllables long and the critical /f/ or /s/ sound was always at the beginning of the third syllable. In addition, taking into account that articulation of fricatives is affected by the vocalic context, we made sure that for all items the critical sound was preceded by an /o/ or /u/ sound (a low vowel) and followed by an /e/ or /i/ sound (a high vowel). Items were checked by a Spanish-Basque bilingual research assistant to make sure that they were nonwords in both Spanish and Basque, but phonotactically and morphologically consistent with Spanish. Each participant was randomly assigned to one of two stimulus groups so that the same item would contain an /f/ sound for half of the participants and an /s/ sound for the other half (e.g., ranpo<u>f</u>ita, ranpo<u>s</u>ita; see Table 3).

Table 3. *List of 12 novel words with an /s/ or /f/ sound per stimulus group*

Group 1	**Group2**
ranpo<u>f</u>íta	ranpo<u>s</u>íta
kenu<u>f</u>éðo	kenu<u>s</u>éðo
miðo<u>f</u>éla	miðo<u>s</u>éla
intu<u>f</u>éna	intu<u>s</u>éna
gaɾto<u>f</u>ípo	gaɾto<u>s</u>ípo
beɲu<u>f</u>íko	beɲu<u>s</u>íko
nulto<u>s</u>éra	nulto<u>f</u>éra
tɾaðo<u>s</u>íta	tɾaðo<u>f</u>íta

[2] We added background noise in the stimuli used in this task in case the versions without noise produced ceiling effects.

erpuséno	erpuféno
ditoséla	ditoféla
liβ̬usíɣ̬o	liβ̬ufíɣ̬o
parpusíno	parpufíno

Auditory stimuli for this task were created using recordings of natural speech spoken by a male[3] native Spanish speaker. After pre-processing, 50 ms of silence was added before and after each word.

Visual stimuli consisted of color pictures of 12 unfamiliar objects, which had been collected online and used as unusual objects in previous experiments (Leach & Samuel, 2007). Images measured 240 × 240 pixels during presentation.

5AFC picture identification task. The 24 real word fillers were a subset of the 33 real word items used in the old/new task of Experiment 1 and the same auditory stimuli were used. These items did not contain any /f/ or /s/ sounds. For the 12 novel words, auditory stimuli were created using recordings of natural speech spoken by the same female speaker as in Experiment 1. To construct the mispronounced stimuli (i.e., containing ambiguous /f-s/ sounds), we followed the same procedures as in Experiment 1.

For the 12 novel words, the same images were used as in training. For each of the 12 visible fillers, a picture was collected online and edited as needed. Images measured 240 × 240 pixels during presentation.

[3] Training and testing stimuli were spoken by two speakers of different sex. This was done because pilot data showed that when the same voice was used for training and testing, no phonemic recalibration was detected. This drop in the effect likely reflects the fact that exposure to unambiguous tokens of critical speech sounds can interfere with phonemic recalibration (Kraljic et al., 2008). In addition, using speakers of different genders in training and testing allowed us to assess the establishment of lexical representations independently of voice-specific information. We come back to these issues in the General Discussion.

2AFC usi/ufi task: We used the stimuli from the corresponding Experiment 1 task.

Word recognition task. Auditory stimuli were created using separate recordings of natural speech spoken by the same female speaker as in the 5AFC task. Three versions of each of the 12 novel words were used (one to use as target, e.g., ranpofíta, and two as lures, e.g., rantofita and ranpofila). The same audio pre-processing steps were used as before. In addition, copies of all stimuli were combined with white Gaussian noise at a zero signal-to-noise ratio (0 SNR) to create the corresponding noisy versions.

Results

One participant did not come back on Day 2. In addition, data from four sessions were excluded from the analyses due to technical errors (one from Day 1 and three from Day 2). This left us with valid data from 148 Day 1 sessions and 145 Day 2 sessions (i.e., 17-19 participants per Training Condition × Exposure × Day combination). Data are openly available at: https://osf.io/7scvw/?view_only=00e85ddaa87b457a8d8c2fd64fcf0aea.

We present the results in six sections. The first three sections include descriptive analyses and document the basic effect of phonemic recalibration in the baseline, Perception-Only condition. The fourth section addresses the primary question of whether immediate and/or delayed production modulates the phonemic recalibration effect. Finally, the last two sections address our two secondary questions regarding the role of consolidation and the effect of immediate and delayed production on lexical configuration, respectively.

Training: RTs and accuracy across conditions

Participants performed the 2AFC task without difficulties and responded in a prompt manner. On average, accuracy on the last (20^{th}) block of training was at 96.6% on Day 1 and at 98.1% on

Day 2 (see Fig. 5). As in previous studies (e.g., Leach & Samuel, 2007; Samuel & Larraza, 2015), performance asymptotes near ceiling after 8-10 training exposures.

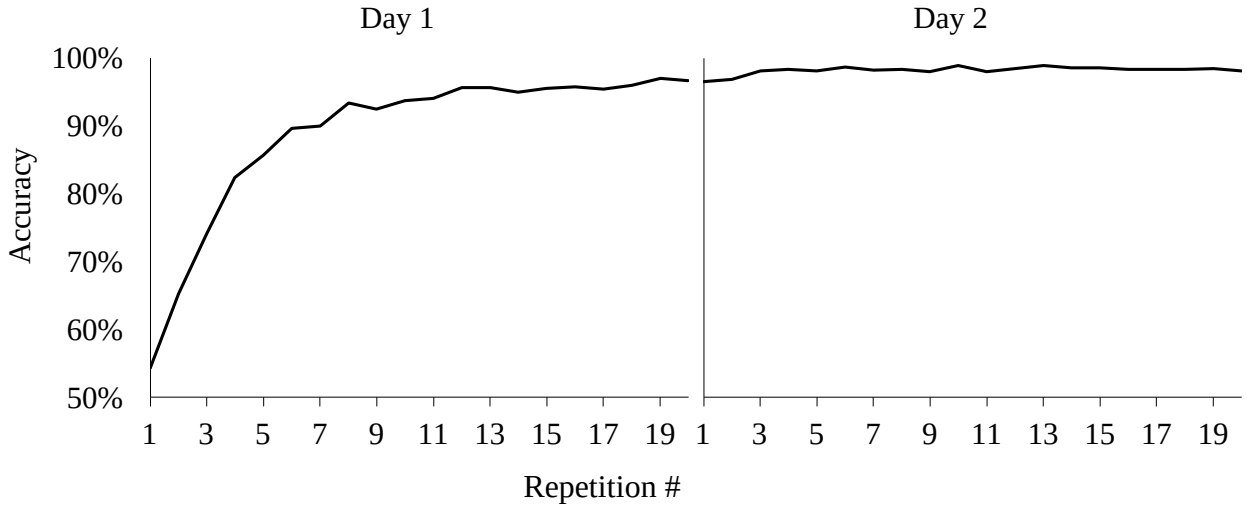


Figure 5. Average accuracy as a function of repetition on Day 1 and Day 2

Participants' vocal 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 and placement of response time (RT) marks. Average accuracy was at 93.7% (SD = 6.4%) and average RT was 417 ms (SD = 166 ms).

Testing: Data inspection across conditions

As in Experiment 1, we calculated the average proportion of "ufi" versus "usi" responses for each /ufi-usi/ step for each participant on Day 1. The difference between the two exposure groups (i.e., participants exposed to ambiguous /f/ vs /s/) was used as a measure of word learning. Figure 6A shows the average proportion of "usi" responses per /ufi-usi/ step for participants exposed to ambiguous /f/ versus ambiguous /s/ sounds.

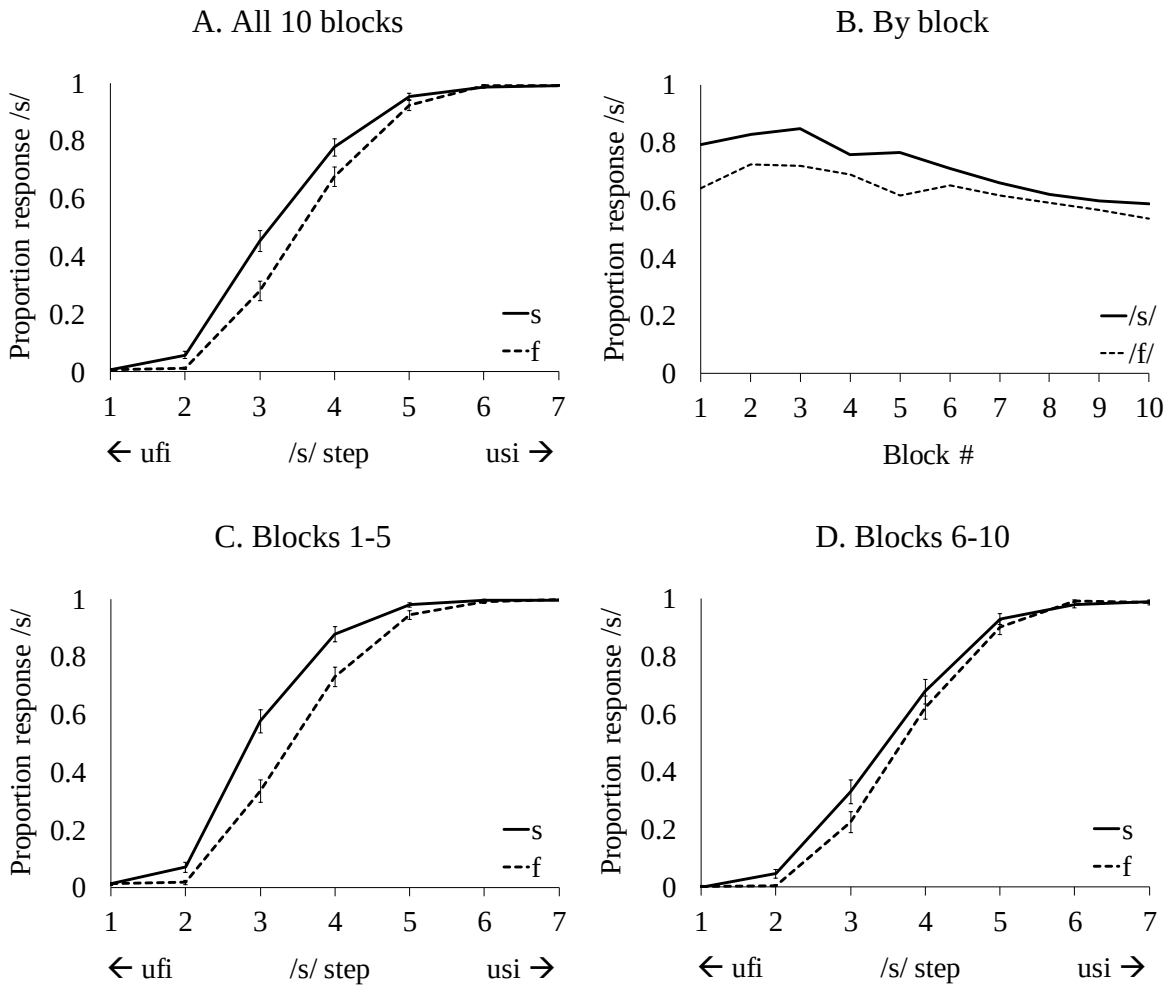


Figure 6. Proportion of /usi/ responses per /ufi-usi/ step by exposure condition (/s/ versus /f/) on Day 1 averaged across training conditions, and testing blocks (panel A). Proportion of /usi/ responses on Day 1 averaged across the three middle /ufi-usi/ steps as a function of exposure condition (/s/ versus /f/) and testing block (panel B). Panels C and D show the Panel A data in the first versus second half of testing. Note: Error bars indicate standard error of the mean.

Overall, the data in Figure 6A look similar to Experiment 1, with participants exposed to ambiguous /s/ sounds being more likely to classify ambiguous /u?i/ items as /usi/. The results replicate Leach and Samuel's (2007) finding that recently-learned words can support phonemic recalibration. That said, we inspected the data split by testing block in order to examine whether the effect was stable during the testing phase. This decision was prompted by recent findings

showing that phonemic recalibration effects tend to decline over the course of the identification test (Caplan et al., 2021; Liu & Jaeger, 2018). In addition, our own initial failed attempt to detect phonemic recalibration when participants were previously exposed –in training– to properly pronounced /s/ and /f/ sounds by the same speaker as in testing (see Footnote 4) suggests the same issue: Hearing the same speaker produce non-ambiguous utterances of the critical sounds seems to destabilize the phonemic recalibration effect. As seen in Figures 6B, 6C, and 6D, our concern was verified: Effects are relatively large early in the testing phase (Figure 6C), but are much smaller later in testing (Figure 6D). As Figures 6B and 6C show, the recalibration effects seem to be relatively stable for the first five passes. As a result, the following data analyses are based on only the first half of the identification test (i.e., blocks 1-5) of the first session.

Regarding the magnitude of the recalibration effect across steps, we see that –as expected– it is the largest in the middle of the continuum (i.e., in the most ambiguous steps; see Figs 6A and 6C). Specifically, the effect appears to be the most robust in steps two (2) through five (5), which is largely in line with our a priori decision to focus our analyses on the middle of the continuum (as in Experiment 1). Taking all things into account, and to ensure that our analyses were optimally sensitive to the specific data, we decided to use these four steps to compute the magnitude of the effect to include in the statistical analyses (for a discussion of the strength of this analysis approach, see Samuel & Dumay, 2021).

Testing: Documenting the basic effect of phonemic recalibration without production

The studies reviewed in the Introduction showed that producing words during the learning phase can disrupt the development of lexical engagement. In order to look at this effect, and to see if it is modulated by the timing of the production component, it is first necessary to establish that, with our stimuli and procedures, words learned without production do in fact support lexical

engagement. Thus, before addressing our main question of how production modulates word learning, we assessed how well novel words that were learned in the baseline, Perception-Only training condition supported phonemic recalibration, our measure of lexical engagement.

The left panel of Figure 7A shows the relevant results. As the figure shows, participants exposed to ambiguous /s/ sounds were more likely to classify ambiguous /u?i/ items as /usi/ than participants who had been exposed to ambiguous /f/ sounds. We tested the recalibration effect (i.e., effect of exposure) using an independent-samples t-test with proportion of /usi/ responses (empirical-logit-transformed) averaged across the middle four steps of the continuum as the dependent variable. For the Perception-Only group, the difference between the two Exposure groups was significant, $t(35)=3.428$, $p=.002$, with participants exposed to ambiguous /s/ sounds being about 17.5% more likely to classify ambiguous /u?i/ items as /usi/ than participants exposed to ambiguous /f/ sounds. These results suggest that new lexical representations that were learned without production were robust enough to drive phonemic recalibration.

Testing: Assessing the effect of immediate and delayed production on word learning

Next, we turn to our main question regarding the effect of immediate and delayed production on novel word learning. Given the pattern reported by Baese-Berk and Samuel (2022) we expected that learning would be the most robust in the Perception-Only condition and the least robust in the Immediate-Production condition, with the two conditions with delayed production being less impaired than the Immediate-Production case (see Figure 1).

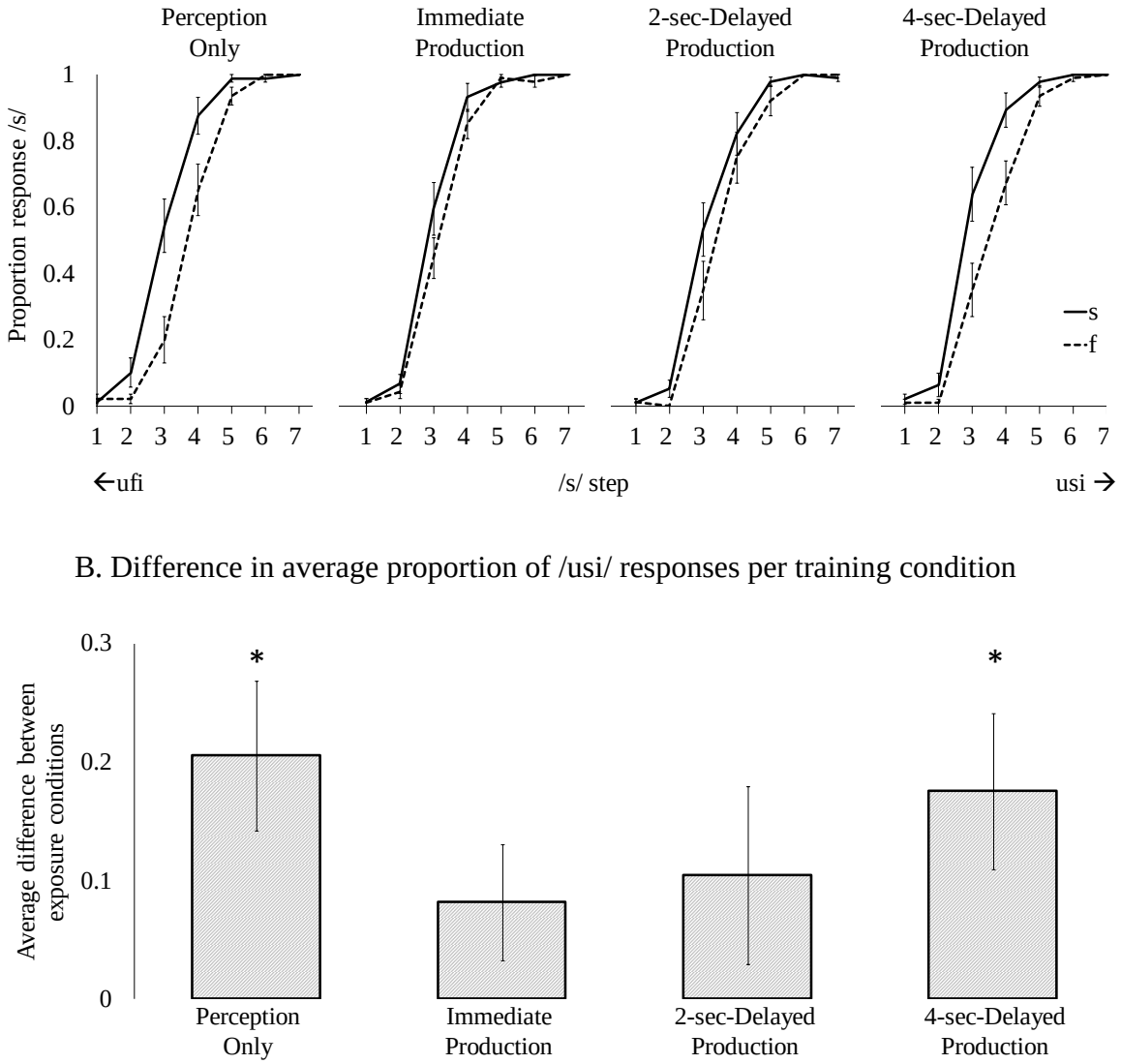


Figure 7. Panel A: Average proportion of /usi/ responses per continuum step by exposure and training condition during the first session of Experiment 2. Panel B: Average differences between exposure conditions in the proportion of /usi/ responses across the middle /ufi-usi/ steps. Note: Error bars indicate standard error of the mean (difference).

Figure 7A shows the average proportion of "usi" responses per step as a function of exposure (ambiguous /f/ versus ambiguous /s/) and training condition on Day 1, while Figure 7B plots the size of this recalibration effect specifically for the four middle steps of the continuum. Across

training conditions, the difference between the two exposure conditions is in the expected direction; participants exposed to ambiguous /s/ sounds were more likely to classify ambiguous /u?i/ items as /usi/. In addition, we see fluctuations in the magnitude of this difference that are in line with our predictions: Perception-Only shows the largest recalibration effect, Immediate-Production shows the smallest one, and the two delayed-Production conditions fall somewhere in-between. Critically, this pattern appears strikingly similar to the pattern reported by Baese-Berk and Samuel (2022; see Figures 1 and 7B).

To statistically assess the effect of each type of production (immediate and with 2 or 4 secs delay), we conducted three 2 (Exposure: ambiguous /f/ vs /s/) × 2 (Training Condition) between-subjects ANOVAs, each contrasting one of the three Production conditions against the Perception-Only (baseline) condition[4]. As before, the dependent variable was proportion of /usi/ responses (empirical-logit-transformed) averaged across the middle four (2-5) steps of the continuum. We expected: (1) a main effect of Exposure (i.e., more "s" responses for the /s/-biased group), (2) a significant Exposure × Training Condition interaction when comparing Immediate-Production against Perception-Only (i.e., larger effect of Exposure for the Perception-Only condition), and (3) a non-significant Exposure × Training Condition interaction when comparing 4-seconds-Delayed-Production against Perception-Only. For the 2-seconds-Delayed-

[4] Given the availability of different statistical approaches, as well as the rising popularity of mixed effects models, it is worth explaining the rationale of our analytical approach. We used ANOVAs instead of a mixed effects model 1) to make the results more directly comparable to previous work (e.g., Leach and Samuel) and 2) because in this case using mixed effects would not likely have an advantage over ANOVA. This is because a) there was only one item (ufi/usi), so there was no item-driven variability to account for, and b) any subject-driven effects would be confounded with condition-driven effects, given that our critical manipulation was between-subjects (that is, each subject was exposed to either ambiguous /s/s or ambiguous /f/s). Moreover, our goal was to test whether each of the three types of production (Immediate/2-secs-Delayed/4-secs-Delayed) led to different degrees of lexical integration when compared to the baseline (Perception-Only) condition. This is why we opted for an analytical approach that would test these specific questions (i.e., three ANOVAs, each testing the corresponding question).

production condition, we expected a pattern similar either to the Immediate- or the 4-seconds-Delayed-Production condition, or something in-between.

The expected main effect of Exposure was significant in all analyses, while the expected Exposure × Training Condition interaction for the Immediate-Production condition was marginally significant, $F(1,69)=2.954$, $p=.090$, $\eta^2 = .041$. In line with our predictions, none of the other interactions was significant (see Appendix Tables A3-A5 for full results). Next, we took a closer look at how word learning outcomes differed among training conditions. We conducted four Bonferroni-corrected comparisons to assess the simple effect of Exposure in each training group (see Figure 7B). As before, the dependent variable was proportion of /usi/ responses (empirical-logit-transformed) averaged across the middle four steps of the continuum. Consistent with prior work (Leach & Samuel, 2007), the simple effect of Exposure was significant for the Perception-Only group, $F(1,140)=11.731$, $p=.001$, $\eta^2=.077$, but not for the Immediate-Production group, $F(1,140)=1.497$, $p=.223$, $\eta^2=.011$. The critical question is whether the new delayed production conditions would pattern with the Perception-Only case (i.e., a significant effect of Exposure), or with the Immediate-Production case (i.e., no significant effect of Exposure). When a delay of two seconds was added before production, the simple effect of Exposure was marginally significant, $F(1,140)=3.296$, $p=.072$, $\eta^2=.023$. Critically, the effect was significant for participants in the 4-seconds-Delayed-Production group, $F(1,140)=8.565$. $p=.004$, $\eta^2=.058$ (see Appendix Table A6 for full results).

In sum, the results show that recalibration is robust when no production is required and becomes small and non-significant with immediate production; a short delay (2 sec) yields marginally significant recalibration, and a longer delay (4 sec) restores the size and significance

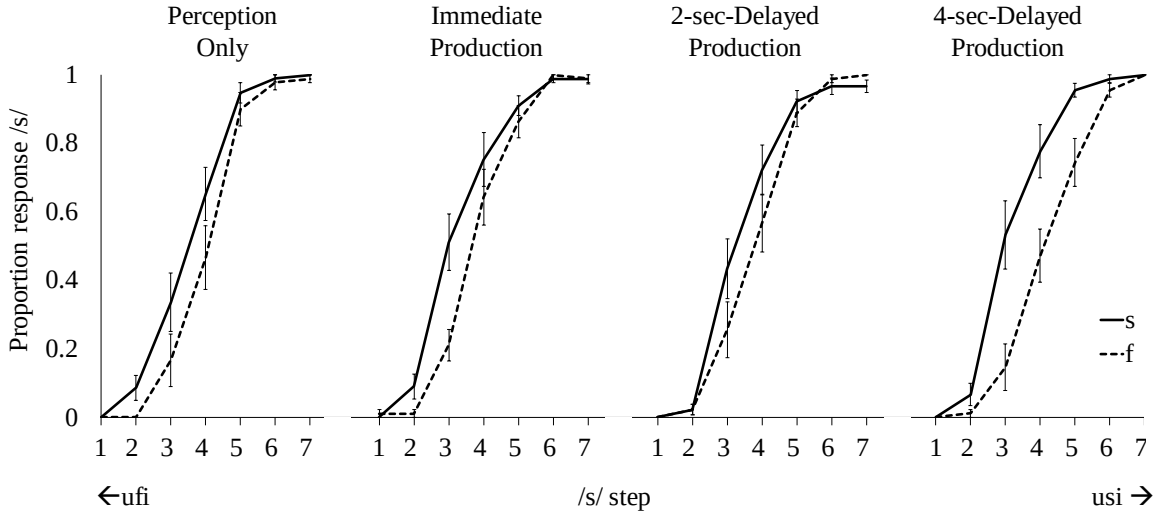
of recalibration. This pattern replicates the results reported by Baese-Berk and Samuel (2022) in their study of learning a new non-native phonemic contrast (see Figures 1 and 7B).

Having addressed our principal question, we next turn to our secondary questions, regarding (a) the possible role of (sleep-driven) consolidation in modulating the effect of production and (b) the effect of production on lexical configuration. To address these, we included data from both days in our analyses (while still only including data from the first five blocks of each identification test).

Testing: Assessing the role of consolidation

Figure 8A plots the proportions of "usi" report for the four conditions on Day 2, and Figure 8B shows the corresponding difference scores for each of the conditions on Day 1 versus Day 2. Overall, we again see that participants exposed to ambiguous /s/ sounds were more likely to classify ambiguous /u?i/ items as /usi/. As on Day 1, the size of these shifts varied across conditions.

A. Average proportion of /usi/ responses per step, exposure, and training condition (Day 2)



B. Difference in average proportion of /usi/ responses per training condition (per Day)

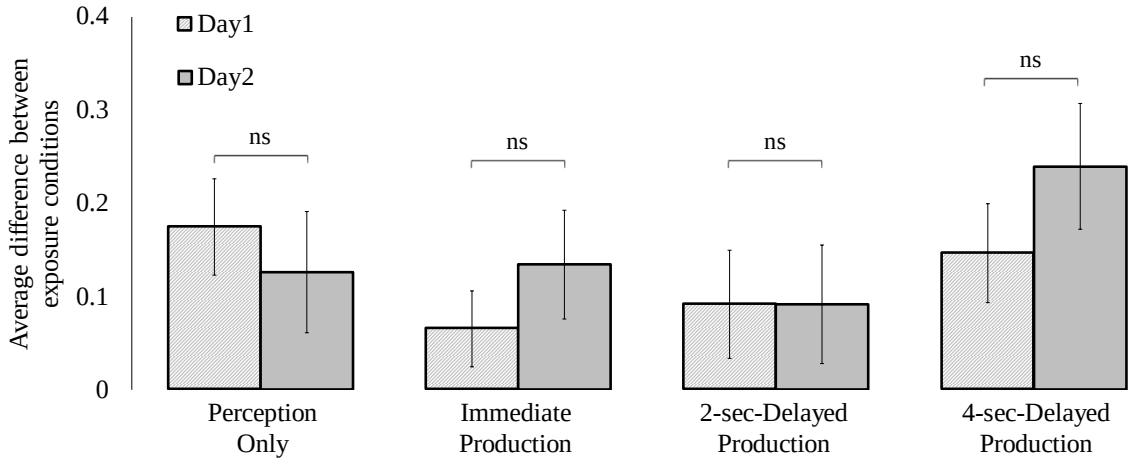


Figure 8. Panel A: Average proportions of /usi/ responses per continuum step by exposure and training condition on the second session of Experiment 2. Panel B: Average differences between exposure conditions in the proportion of /usi/ responses across the middle /ufi-usi/ steps on Day 1 and Day 2. Note: Error bars indicate standard error of the mean (difference).

To statistically assess whether the effects of production differed between days, we followed the same analytical approach as before, but added Day as an independent variable; that is, we conducted three 2 (Exposure: ambiguous /f/ vs /s/) × 2 (Training Condition) × 2 (Day 1 vs. Day 2) repeated-measures ANOVAs, each contrasting one of the three Production conditions against the Perception-Only (baseline) condition. The same dependent variable was used as before (proportion of /usi/ responses, empirical-logit-transformed). If consolidation modulates the detrimental effect of (immediate/delayed) production, this modulation would be reflected in a significant 3-way interaction when contrasting the corresponding production condition against the Perception-Only (baseline) condition. This did not occur: None of the three-way interactions was significant, pointing to the relative stability of the observed Day 1 pattern across days.

However, it is noteworthy that the 3-way interaction was marginally significant for the Immediate-Production, $F(1,68)=3.823$, $p=.055$, $\eta^2=.053$, and the 4-seconds-Delayed-Production

condition, $F(1,68)=3.873$, $p=.053$, $\eta^2=.054$ (see Appendix Tables A7-A9 for full results). Though not significant, these outcomes reflect small changes in opposite directions from Day 1 to Day 2 for the Perception-Only versus these two Production conditions. First, as seen in Figure 8B, there is a small drop in the Exposure effect for the Perception-Only condition. We conducted a two-way Exposure × Day ANOVA to statistically assess this change, which yielded a non-significant interaction, $p=.202$. Second, Figure 8B also shows a small increase from Day 1 to Day 2 for the Immediate-Production and the 4-seconds-Delayed-Production groups. Again, these changes were assessed with two-way ANOVAS and neither of the 2-way interactions was significant, $p=.153$, $p=.149$. Although not significant, these points carry some theoretical interest, which is why we will come back to them in the General Discussion.

Testing: Assessing lexical configuration

We now turn to our last question: Is there an effect of immediate or delayed production on lexical configuration? Recall that this was assessed using a word recognition task administered at the very end of the second session. Participants performed the task without problems, at an accuracy of 72% for the stimuli presented first in noise, and 83% for the stimuli presented second without noise.

Given the nature of the task (i.e., to correctly identify each stimulus as a previously learned word vs. a lure), we calculated a d' prime score for each participant and each stimulus type (with/without noise). Based on previous results, we expected to find a facilitatory effect of production. To assess this, we split the data by training condition (see Fig. 9).

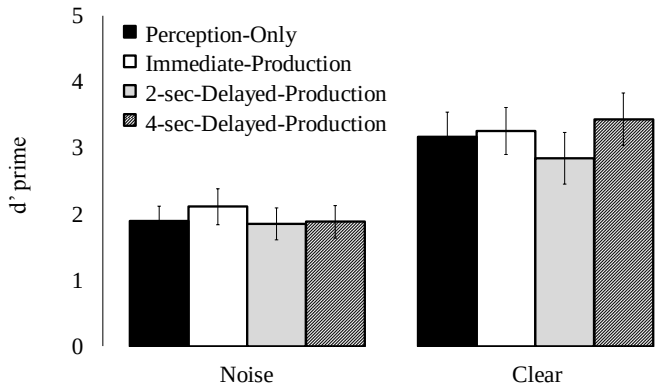


Figure 9. Average d' prime score in word recognition task per training condition. Note: Error bars indicate standard error of the mean.

As seen on Figure 9, no notable differences appear other than the main effect of presentation (first/with noise vs. second/without noise). Although the difference between the Immediate-Production case and the Perception-Only case is in the direction reported by Leach and Samuel (2007), the difference is quite small. To statistically assess whether the results differed between training conditions, we conducted three 2 (Presentation Mode: with vs without noise) × 2 (Training Condition) repeated-measures ANOVAs, each contrasting one of the three Production conditions with the Perception-Only (baseline) condition. In all cases, there was a main effect of Presentation Mode indicating better performance when the words were presented without noise. No other main effects or interactions were significant (see Appendix Tables A10-A12 for full results).

Discussion

The Perception-Only condition served as our baseline to look for effects of production, and the results for this baseline condition aligned with those of Experiment 1. That is, systematic exposure to ambiguous /f-s/ speech sounds led to lexically-driven phonemic recalibration. Participants exposed to mispronounced /f/ sounds learned to perceive ambiguous /f-s/ sounds as

/f/, while participants exposed to mispronounced /s/ sounds learned to perceive them as /s/. Importantly, the items in which ambiguous /f-s/ sounds were embedded during the exposure phase were newly learned words rather than the familiar words used in Experiment 1. This result demonstrates that newly learned words can be incorporated into the mental lexicon well enough to drive perceptual recalibration. Importantly, as expected, having participants repeat each new word immediately after hearing it (i.e., Immediate-Production condition) eliminated the ability of the words to drive the phonemic recalibration effect. The results from these two conditions replicate the core finding of Leach and Samuel (2007) – production during the learning of new words can disrupt the development of their ability to engage with sublexical and lexical representations. This disruption fits with several other studies that have reported detrimental effects of (immediate) production on word learning (e.g., Kapnoula & Samuel, 2022; Kaushanskaya & Yoo, 2011) and learning new phonetic contrasts (e.g., Baese-Berk, 2019; Baese-Berk & Samuel, 2016, 2022).

Our central question was whether inserting a delay between hearing and producing a new word would modulate the detrimental effect of production on word learning. Indeed, when production was delayed, evidence for perceptual recalibration re-surfaced (Fig. 7B). With a 2-second delay, the newly-learned words produced marginally significant recalibration, and with a longer (4-second) delay, the recalibration was robust and reliable.

Testing on Day 2 allowed us to address two secondary questions. Specifically, our findings suggest that (a) the Day 1 pattern is relatively stable across days, and (b) production may not affect lexical configuration in the same way as it affects lexical engagement (Fig. 9). The different patterns of lexically-driven recalibration for the different exposure conditions were not mirrored in the simple recognition test of lexical configuration. That said, any conclusions drawn

from this comparison should be taken with a grain of salt given that for the lexical configuration test we only have data from Day 2.

General Discussion

The main goal of this study was to assess whether delaying production mitigates its negative effect on novel word learning. To do so, we compared word learning outcomes when training included immediate, delayed (two seconds or four seconds), or no production at all. Our results replicate previous work showing that immediate production has a detrimental effect on word learning. Critically, our results also provide clear evidence that delaying production can eliminate the detrimental effect – following a strikingly similar pattern to perceptual learning reported by Baese-Berk and Samuel (2022; see Figures 1 and 7B). In the following sections, we discuss the potential mechanisms behind this pattern, as well as broader theoretical issues.

On the mechanisms underlying the detrimental effect of production on learning

In the Introduction, we listed three potential mechanisms that could lead to a learning deficit when training includes production: (1) learners may mispronounce new items and these mispronunciations may act as incorrect input, (2) activation of production-related representations may interfere with the learning process, and (3) production may act as a distractor task taking up attentional resources that would otherwise be allocated to perceptual processing.

In our experience, when people find out about the negative effect of production on perceptual learning of lexical or sublexical representations, the first explanation is usually their preferred account. One prior result argued against this interpretation: Baese-Berk (2019) found no relationship between production accuracy and perceptual learning in her study of American English speakers learning a prevoiced-voiced stop distinction; the first account would predict

that more accurate producers would be better perceptual learners because their self-input is more accurate. Our results provide two new important reasons to doubt the first explanation. First, production accuracy during training was generally near ceiling (~94%), and even the small number of errors rarely if ever involved the critical /f/ or /s/ sound. Thus, the participants in the Immediate-Production condition rarely if ever provided incorrect self-input, yet their lexical representations could not support recalibration. Second, participants in the 4-seconds-Delayed-Production condition would presumably be subject to the same "bad" self-generated input as those in the Immediate-Production condition, but there was no detrimental effect of production in this case. Thus, the results of the current study allow us to reject the most intuitive explanation as playing much of a role (if any) in the phenomenon.

Critically, our results provide good support for the second and/or third possibility. As seen in Fig.7B, separating the production requirement from hearing the new word by two seconds partly alleviated the detrimental effect of production, while adding a longer separation (four seconds) eliminated it. This pattern suggests that right after hearing a new word, listeners need a moment to process the input; certain computations may be needed to consolidate the trial's learning. If production is required during that moment, it disrupts this process. The second explanation would attribute this disruption to an incompatibility of the developing perceptual encoding with the activated production representation (i.e., there is some form of active competition). Alternatively, the third explanation is that production could hurt perceptual learning because it distracts listeners from processing the perceptual input during that critical moment. Baese-Berk and Samuel (2016, 2022) have provided evidence that supports both of these potential mechanisms for learning at the segmental level (see also Wright, 2021). They found that a requirement to produce something other than the to-be-learned sound (i.e., naming aloud a

printed letter, rather than the syllable heard during the training trial) led to significantly worse perceptual learning than having no production requirement, but to significantly better perceptual learning than saying the matching syllable. The first of these two differences is predicted by the distraction account (Explanation 3), while the second difference aligns with the more specific incompatibility account (Explanation 2).

It is informative that no effect of training condition was found in the word recognition task, which measured the lexical configuration of new items – i.e., a property roughly corresponding to the initial encoding of new words into the mental lexicon. This suggests that the detrimental effect of production is specific to lexical engagement, which refers to processing-related aspects of a lexical entry. Our results are in partial agreement with Leach and Samuel (2007), who also assessed the effect of production on each of the two properties and, similarly to the present study, only found a detrimental effect on lexical engagement (in that study, there was even a positive effect of production on lexical configuration, despite the disruption of lexical engagement). Converging evidence comes from Kapnoula and Samuel, (2022), who examined the effect of production as a function of time (both at a millisecond scale and with respect to early versus late training trials). In that study, production had a weak facilitatory effect in the initial encoding of new words, but later its effect became detrimental. Specifically, using eye movements to a picture of a word's referent as a proxy for lexical activation, there were shallower activation slopes for words that were learned with production than without it. This pattern suggests that production prevented those words from becoming well integrated into the system and, as a result, their recognition did not reach the same level of automatization as that of words that were only heard during training.

Taken together, the results suggest that including an immediate production requirement during word learning prevents new items from becoming well integrated into the mental lexicon. However, if perception and production are separated by a long-enough delay (~four seconds), then the two processes do not have to compete for the same attentional resources (and/or, incompatible representations are not simultaneously active) and, thus, production does not interfere with processing of the perceptual input.

How is the effect of production modulated by overnight consolidation?

In addition to this core question, we also examined how the effect of production is modulated by consolidation. To look at that, we repeated the training and testing sessions 24 hours later. Our results revealed that the effect of Exposure was not significantly different across days for any of the training conditions.

Even though none of the 3-way interactions was significant, two were marginally significant, reflecting a contrasting pattern in how the recalibration effect changed across days in the Perception-Only versus two of the Production conditions. Follow-up analyses showed that these changes were not significant; however, we think that this pattern merits some speculation, and investigation in future work. First, there was a small (non-significant) decrease in the Exposure effect from Day 1 to Day 2 for items in the Perception-Only condition, which we think might trace to the recent finding that phonemic recalibration diminishes when listeners are exposed to test stimuli (Caplan et al., 2021; Liu & Jaeger, 2018). That is, this trend may be a continuation of the decreasing trend that was observed within the first session of Day 1 (see Fig. 6B).

Second, we observed small (non-significant) increases of the Exposure effect in the Immediate- and the 4-seconds-Delayed-Production conditions. Our speculation is that production

leads to increased input variability, which in turn can boost consolidation of word learning. That is, each time a participant produced a word, they naturally said it in a slightly different way. These acoustically variable utterances then acted as perceptual input because participants hear their own productions. Previous research suggests that higher acoustic variability in irrelevant dimensions helps listeners identify the dimensions that are relevant, thus boosting learning (Rost & McMurray, 2009, 2010). Therefore, it is possible that production can help learning by increasing input variability, leading to better stabilized lexical representations and preventing the fading of perceptual recalibration observed in the Perception-Only condition. Prior research indicates that this kind of lexical stabilization occurs overnight. There is growing evidence that sleep-driven consolidation boosts word learning (Bakker et al., 2014; Dumay & Gaskell, 2007; Gaskell & Dumay, 2003; Tamminen et al., 2010, see McMurray et al., 2016, and Palma & Titone, 2020, for reviews). Sleep not only consolidates the memory traces of newly learned words, it may also strengthen the connections among diverse utterances of the same word, leading to a better inter-connected and thus more robust lexical representation. The growth from Day 1 to Day 2 in this condition is consistent with this.

As we noted, when focusing on each of the training conditions, none of the changes across days was significant; however, when comparing the Production conditions against the Perception-Only baseline, marginal statistical outcomes emerge at the level of 3-way interactions in two out of the three comparisons (for both, $.05 < p < .06$). Taking a step back and looking at the bigger picture, it seems that the declining tendency observed in the Perception-Only condition is not only absent in the two Production conditions, the pattern is in fact reversed. This suggests that there is some production-driven process that strengthens the recalibration effect

overnight, overlaid on any decrease in the effect due to participants' exposure to the test stimuli. This pattern could be a fruitful direction for future investigation.

Concluding considerations

We found that inserting a four second delay before production alleviated its detrimental effect on word learning, and even led to improved performance on the next day. Note that adding this delay creates a situation in which presentation of the to-be-learned items is more spread out over time, compared to the Immediate-Production condition. Thus, better learning in the 4-seconds-Delayed-Production condition is consistent with the well-established *spacing effect*, according to which learning is improved when items are temporally more widely spaced compared to when they are presented closely together (Carpenter et al., 2012; Kelley & Whatson, 2013). That said, the spacing effect would only potentially apply to the differences among the three production conditions, but it makes the incorrect prediction that the Immediate-Production condition should lead to better learning than the Perception-Only condition – the results are the opposite.

Our results add to the set of studies reporting detrimental effects of production on word learning (Kapnoula & Samuel, 2022; Leach & Samuel, 2007; Zamuner et al., 2018). In addition, our findings show that adding a delay before the production requirement alleviates its detrimental effect. This pattern provides novel insights into the cognitive mechanisms underlying this effect. Asking learners to repeat new words immediately after hearing them can distract them from fully processing the speech input and may lead to simultaneous activation of incompatible perceptual and production representations. As a result, perceptual learning of the words is hindered. The fact that the same pattern has been found when listeners are learning a

non-native phonetic contrast (Baese-Berk & Samuel, 2022) indicates that this is a general property of the language system.

There is now a growing body of work pointing to the negative role that production can play in different aspects of language-related learning (Baese-Berk, 2019; Baese-Berk & Samuel, 2016, 2022; Kapnoula & Samuel, 2022; Kaushanskaya & Yoo, 2011; Wright, 2021). These results complement the more intuitive finding that production can help learning when it is incorporated into training in a way that does not disrupt perceptual processing.

Finally, in addition to the theoretical value of these results in pinpointing the mechanism via which production disrupts novel word learning, there is also clear potential for real-world applications. For example, our findings strongly suggest that second language learning practices could be optimized by incorporating pauses (at the right places) during vocabulary exercises in which a teacher models words for the students to produce so that students have enough time to perceptually process words that they are learning.

Acknowledgements

The authors thank Alexandra Kypta-Vivanco for help with data collection. Support for this project was provided by the Spanish Ministry of Science and Innovation through Grant # PSI2017-82563-P, awarded to AGS, and Grant # PID2020-113348GB-I00, awarded to AGS and ECK, 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 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

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. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language, 89*, 23–36. https://doi.org/10.1016/J.JML.2015.10.008

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

Bakker, I., Takashima, A., van Hell, J. G., Janzen, G., & McQueen, J. M. (2014). Competition from unseen or unheard novel words: Lexical consolidation across modalities. *Journal of Memory and Language, 73*, 116–130. https://doi.org/10.1016/j.jml.2014.03.002

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual Recalibration of Auditory Speech Identification: A McGurk Aftereffect. *Psychological Science*, *14*(6), 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Boersma, P., & Weenink, D. (2016). *Praat: doing phonetics by computer [Computer program]*.

Caplan, S., Hafri, A., & Trueswell, J. C. (2021). Now You Hear Me, Later You Don't: The Immediacy of Linguistic Computation and the Representation of Speech. *Psychological Science*, *32*(3). https://doi.org/10.1177/0956797620968787

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H. K., & Pashler, H. (2012). Using Spacing to Enhance Diverse Forms of Learning: Review of Recent Research and Implications for

Instruction. In *Educational Psychology Review* (Vol. 24, Issue 3, pp. 369–378). Springer. https://doi.org/10.1007/s10648-012-9205-z

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*(3), 341–361.

Darwin, C. J., Turvey, M. T., & Crowder, R. G. (1972). An Auditory Analogue of the Sperling Partial Report Procedure: Evidence for Brief Auditory Storage1. In *COGNITIVE PSYCHOLOGY* (Vol. 3).

Davis, M., & Gaskell, M. G. (2009). A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1536), 3773–3800. https://doi.org/10.1098/rstb.2009.0111

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35–39.

Dumay, N., & Gaskell, M. G. (2012). Overnight lexical consolidation revealed by speech segmentation. *Cognition*, *123*(1), 119–132. https://doi.org/10.1016/j.cognition.2011.12.009

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., & 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

Kapnoula, E. C., & McMurray, B. (2016). Newly learned word forms are abstract and integrated

immediately after acquisition. *Psychonomic Bulletin and Review*, *23*(2). 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. (2022). Reconciling the contradictory effects of production on word learning: Production helps at first, but hurts later. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *48*(3), 394–415.

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

Kelley, P., & Whatson, T. (2013). Making long-term memories in minutes: a spaced learning pattern from memory research in education. *Frontiers in Human Neuroscience*, *7*(SEP), 589. https://doi.org/10.3389/fnhum.2013.00589

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability: Research article. *Psychological Science*, *19*(4), 332–338. https://doi.org/10.1111/j.1467-9280.2008.02090.x

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

Liu, L., & Jaeger, T. F. (2018). Inferring causes during speech perception. *Cognition*, *174*, 55–70. https://doi.org/10.1016/j.cognition.2018.01.003

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

Mama, Y., & Icht, M. (2016). Auditioning the distinctiveness account: Expanding the production effect to the auditory modality reveals the superiority of writing over vocalising. *Memory*, *24*(1), 98–113. https://doi.org/10.1080/09658211.2014.986135

Mama, Y., & Icht, M. (2018). Production on hold: delaying vocal production enhances the production effect in free recall. *Memory*, *26*(5), 589–602. https://doi.org/10.1080/09658211.2017.1384496

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

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-11. https://doi.org/10.1016/s0010-0277(01)00157-3

McMurray, B., Kapnoula, 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, pp. 126–152). https://doi.org/10.4324/9781315772110

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

Palma, P., & Titone, D. (2020). Something old, something new: A review of the literature on

sleep-related lexicalization of novel words in adults. *Psychonomic Bulletin & Review*, 1–26. https://doi.org/10.3758/s13423-020-01809-5

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

Samuel, A. G. (1986). *Red herring detectors and speech perception: In defense of selective adaptation*. *18*(4), 452–499. https://doi.org/10.1016/0010-0285(86)90007-1

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics*, *71*(6), 1207–1218. https://doi.org/10.3758/APP.71.6.1207

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

Sarrett, M. E., McMurray, B., & Kapnoula, E. C. (2020). Dynamic EEG analysis during language comprehension reveals interactive cascades between perceptual processing and sentential expectations. *Brain and Language*, *211*, 104875. https://doi.org/10.1016/j.bandl.2020.104875

Tamminen, J., Payne, J. D., Stickgold, R., Wamsley, E. J., & Gaskell, M. G. (2010). Sleep spindle activity is associated with the integration of new memories and existing knowledge. *The Journal of Neuroscience*, *30*(43), 14356–14360.

https://doi.org/10.1523/JNEUROSCI.3028-10.2010

Wright, J. (2021). *Factors Affecting the Incidental Formation of Novel Suprasegmental Categories*. University of Oregon.

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. *List of filler items used in the exposure phase (old/new task) of Experiment 1*

"Old" stimuli		"New" stimuli	
Real words	Nonwords	Real words	Nonwords
abanico [=hand fan]	anapriga	agujero [=hole]	anutergo
bocadillo [=sandwich]	deribura	antidoto [=antidote]	binruhete
dinamita [=dynamite]	gartidalo	ingratitude [=ingratitude]	bilojidor
hermanito [=little brother]	imanpigo	batallador [=warrior]	dolitaro
lagartija [=lizard]	konkoripa	detonante [=explosive]	emagride
molinero [=miller]	morilajo	caramelo [=candy]	enimorgo
ovejero [=shepherd]	ondumila	colorante [=dye]	gupirene
reportaje [=reportage]	regamila	grabadora [=tape recorder]	ikanena
unicornio [=unicorn]	urtarija	electrodo [=electrode]	jekubera
		guarderia [=kindergarten]	kitatrupo
		joyeria [=jewlry]	lamaruna
		jugadora [=player]	lertanito
		levadura [=yeast]	mendadila
		maletero [=trunk]	naribanda
		mermelada [=marmelade]	natrajitad
		navarrete [=last name]	olupiro
		neutralidad [=neutrality]	pamentijad
		opereta [=operetta]	prodijunde
		pelicula [=movie]	rantajika
		picadura [=bite]	tanirador
		remolino [=swirl]	trekalidad
		terminator [=terminator]	ugakibo
		termometro [=thermometer]	jendalico
		tonalidad [=tonality]	gimpiriye

Table A2. *List items per set used in the exposure phase (5AFC task) of Experiment 2*

Set #	Novel word #1	Novel word #2	Visible filler #1	Visible filler #2	Invisible filler #1	Invisible filler #2
1	ranpo[f/s]íta	nulto[f/s]éra	mermelada	unicornio	opereta	picadura
2	kenu[f/s]éðo	traðo[f/s]íta	bocadillo	abanico	guarderia	molinero
3	miðo[f/s]éla	liβu[f/s]íɣo	grabadora	termometro	maletero	batallador
4	intu[f/s]éna	paɾpu[f/s]íno	remolino	caramelo	ovejero	levadura
5	gaɾto[f/s]ípo	eɾpu[f/s]éno	dinamita	agujero	pelicula	colorante
6	beɲu[f/s]íko	beɲu[f/s]íko	joyeria	lagartija	electrodo	antidoto

Table A3. *Full 2×2 ANOVA results: Exposure effect (indexing lexical integration) in Immediate-Production condition versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Exposure	0.205	1	0.205	13.463	<.001	0.163
Condition	0.066	1	0.066	4.327	.041	0.059
Exposure × Condition	0.045	1	0.045	2.954	.090	0.041
Error		69				

Table A4. *Full 2×2 ANOVA results: Exposure effect (indexing lexical integration) in 2-seconds-Delayed-Production condition versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Exposure	0.263	1	0.263	12.008	.001	0.146
Condition	0.001	1	0.001	0.520	.821	0.001
Exposure × Condition	0.025	1	0.025	1.133	.291	0.016
Error		70				

Table A5. *Full 2×2 ANOVA results: Exposure effect (indexing lexical integration) in 4-seconds-Delayed-Production condition versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Exposure	0.387	1	0.387	19.160	<.001	0.213
Condition	0.008	1	0.008	0.396	.531	0.006
Exposure × Condition	0.003	1	0.003	0.139	.710	0.002
Error		71				

Table A6. *Full results of post-hoc comparisons (Bonferroni-corrected) between exposure conditions (/s/-bias minus /f/-bias)*

Condition	Difference between exposure conditions (/s/-bias minus /f/-bias	*SE*	*p*	Lower Bound	Upper Bound
Perception-Only	0.156	0.046	.001	0.066	0.246
Immediate-Production condition	0.056	0.046	.223	-0.035	0.148
2-seconds-Delayed-Production condition	0.083	0.046	.072	-0.007	0.173
4-seconds-Delayed-Production condition	0.131	0.045	.004	0.043	0.220

Table A7. *Full 2×2×2 ANOVA results: Immediate-Production versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Exposure	0.391	1	0.391	11.211	.001	0.142

Condition	0.102	1	0.102	2.913	.092	0.041
Day	0.352	1	0.352	50.682	<.001	0.428
Exposure × Condition	0.010	1	0.010	0.275	.602	0.004
Exposure × Day	0.001	1	0.001	0.099	.754	0.001
Condition × Day	<0.001	1	<0.001	0.003	.957	<.001
Exposure × Condition × Day	0.026	1	0.026	3.823	.055	0.053
Error		68				

Table A8. *Full 2×2×2 ANOVA results: 2-seconds-Delayed-Production versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Exposure	0.383	1	0.383	8.307	.005	0.109
Condition	0.016	1	0.016	0.338	.563	0.005
Day	0.244	1	0.244	51.813	<.001	0.432
Exposure × Condition	0.011	1	0.011	0.236	.629	0.003
Exposure × Day	0.006	1	0.006	1.370	.246	0.020
Condition × Day	0.009	1	0.009	1.881	.175	0.027
Exposure × Condition × Day	0.003	1	0.003	0.667	.417	0.010
Error		68				

Table A9. *Full 2×2×2 ANOVA results: 4-seconds-Delayed-Production versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Exposure	0.840	1	0.840	19.260	<.001	0.221
Condition	0.011	1	0.011	0.244	.623	0.004

Day	0.315	1	0.315	45.961	<.001	0.403
Exposure × Condition	0.037	1	0.037	0.857	.358	0.012
Exposure × Day	0.001	1	0.001	0.101	.751	0.001
Condition × Day	0.001	1	0.001	0.112	.739	0.002
Exposure × Condition × Day	0.027	1	0.027	3.873	.053	0.054
Error		68				

Table A10. *Full 2×2 ANOVA results: Recognition d' prime in Immediate-Production condition versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Training Condition	0.862	1	0.862	0.236	.628	0.002
Presentation Mode	53.645	1	53.645	14.701	<.001	0.093
Training Condition × Presentation Mode	0.138	1	0.138	0.038	.846	<.001
Error		144				

Table A11. *Full 2×2 ANOVA results: Recognition d' prime in 2-seconds-Delayed-Production condition versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Training Condition	1.248	1	1.248	0.331	.566	0.002
Presentation Mode	47.813	1	47.813	12.694	<.001	0.080
Training Condition × Presentation Mode	0.696	1	0.696	0.185	.668	<.001
Error		146				

Table A12. *Full 2×2 ANOVA results: Recognition d' prime in 4-seconds-Delayed-Production condition versus baseline (Perception-Only)*

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
Training Condition	0.599	1	0.599	0.156	.693	0.001
Presentation Mode	74.435	1	74.435	19.431	<.001	0.117
Training Condition × Presentation Mode	0.771	1	0.771	0.201	.564	0.001
Error		146				