Producing During Language Learning Improves Comprehension

Elise W. M. Hopman and Maryellen C. MacDonald

University of Wisconsin-Madison

Author Note

Elise W. M. Hopman and Maryellen C. MacDonald, Department of Psychology, University of Wisconsin-Madison.

We thank Teresa Turco for creating stimuli. We thank the Language and Cognitive Neuroscience Lab, Jenny Saffran and Tim Rogers for helpful discussion. We thank Sean Kang and two anonymous reviewers for their valuable comments. Support for this research was provided by the Graduate School and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation.

Correspondence concerning this article should be addressed to Elise W.M. Hopman, Department of Psychology, 618 Brogden Hall, 1202 West Johnson Street, Madison, WI 53706. Contact: hopman@wisc.edu

Abstract

Language learners often spend more time comprehending than producing a new language. However, memory research suggests reasons to suspect that production practice might provide a stronger learning experience than comprehension practice. We tested the benefits of production during language learning and the degree to which this learning transfers to comprehension skill. We taught participants an artificial language containing multiple linguistic dependencies. Participants were randomly assigned to either a production or a comprehension learning condition, with conditions designed to balance attention demands and other known production-comprehension differences. After training, production-learning participants outperformed comprehension-learning participants on vocabulary comprehension and on comprehension tests of grammatical dependencies, even when controlling for individual differences in vocabulary learning. This result shows that producing a language during learning can improve subsequent comprehension, which has implications for theories of memory and learning, language representations, and educational practices.

Keywords: language production, language learning, language comprehension, artificial language learning, learning transfer

Producing During Language Learning Improves Comprehension

Imagine the first day of a foreign language course requiring students to speak in the new language immediately. In this curriculum, students don't simply repeat words but must generate whole grammatical sentences within the first hour. Of course this intense production experience should improve production abilities. More surprising is our hypothesis that production practice will yield improved vocabulary and grammatical *comprehension* abilities compared to students in an intense comprehension-focused curriculum. If so, such results would have implications for language pedagogy as well as theories of memory, learning transfer, and language representations.

Views of the relationship between production and comprehension, and thus the potential for transfer among tasks, range from full involvement of production in comprehension tasks (Pickering & Garrod, 2013) to minimally overlapping systems (Grodzinsky, 2000). More generally, research on learning transfer in nonlinguistic domains shows that some learning does not transfer to new task demands, even with identical materials (Green, Kattner, Siegel, Kersten, & Schrater, 2015). Language comprehension is known to affect native language production (Bock, Dell, Chang, & Onishi, 2007; Montag & MacDonald, 2015), but evidence for the reverse is scant and conflicting (Branigan, Pickering, & McLean, 2005; Segaert, Menenti, Weber, Petersson, & Hagoort, 2011). The influential input hypothesis in second language acquisition (Krashen, 2003) claims that language production practice does not benefit language learning, and related research finds comprehension practice improves second language production but not vice versa (VanPatten & Cadierno, 1993; VanPatten, 2013). Some studies suggest that speech production practice can impair perception (Baese-Berk & Samuel, 2016; Leach & Samuel, 2007), while other similar studies show benefits (Bixby, 2017). Taken together, these findings

suggest that comprehension skill would be best developed with comprehension training, not production training.

In contrast, memory researchers have found that production can boost some types of learning. Production may improve learning in several different ways. First, language production provides the opportunity to both produce and then hear one's own speech, providing both an additional presentation of the material and an alternative modality for encoding it (MacLeod & Bodner, 2017). Second, language production is more attention-demanding than comprehension (Boiteau, Malone, Peters, & Almor, 2014), potentially yielding greater depth of processing, with consequent memory benefits (Craik & Tulving, 1975). Relatedly, production necessitates choices of what to say, and making task-relevant choices improves learning (Carter & Ste-Marie, 2017). Third, comprehension involves recognizing a stimulus, whereas production involves recall from memory, which benefits information retention—the testing effect (Roediger & Karpicke, 2006). In fact, there is even some evidence for learning transfer from recall (production) to recognition (comprehension; Wenger, Thompson, & Bartling, 1980). Furthermore, retrieval practice can guide learning by changing subsequent long-term memory representations (Fan & Turk-Browne, 2013).

These inherent differences between production and comprehension suggest that production experience should improve language learning. Karpicke and Roediger (2008) found that participants learning Swahili-English word pairs benefited significantly more from repeated retrieval (recall) practice than repeated studying (recognition). However, it is unknown whether production benefits extend beyond vocabulary learning. For example, the sentence "That dog with spots runs" conveys its meaning via grammatical elements, word order and number agreement. Both *dog* and *that* are singular, and *dog* agrees with *runs,* even though this noun and

verb are non-adjacent. These features reflect both the hierarchical and sequential structure of languages, and we predict that language production is likely a strong learning tool here (Fig. 1). Beyond the learning benefits described above, production involves planning the serial order of words, which engages serial ordering mechanisms well known in working memory tasks (MacDonald, 2016). Comprehension is more variable: it may be sometimes include careful syntactic analysis but often is "good enough," giving limited attention to syntax and relying on other cues to meaning (Ferreira & Patson, 2007). This shallower processing may provide a poor learning environment for syntactic dependencies compared to sequential processing inherent in language production.

The current study investigates the potential benefits of production in a between-subjects manipulation of learning task (comprehension or production), followed by comprehension tests. Our tasks minimize some well-known production-comprehension learning differences, allowing a focus on inherent processing differences (Fig. 1). Participants learned an artificial language incorporating a strict word order, complex word morphology, and grammatical dependencies across words. We hypothesize that production experience, compared to comprehension experience, will yield improved vocabulary comprehension. Moreover, we hypothesize that the production learning group will have improved comprehension of grammatical dependencies, even when controlling for vocabulary comprehension.

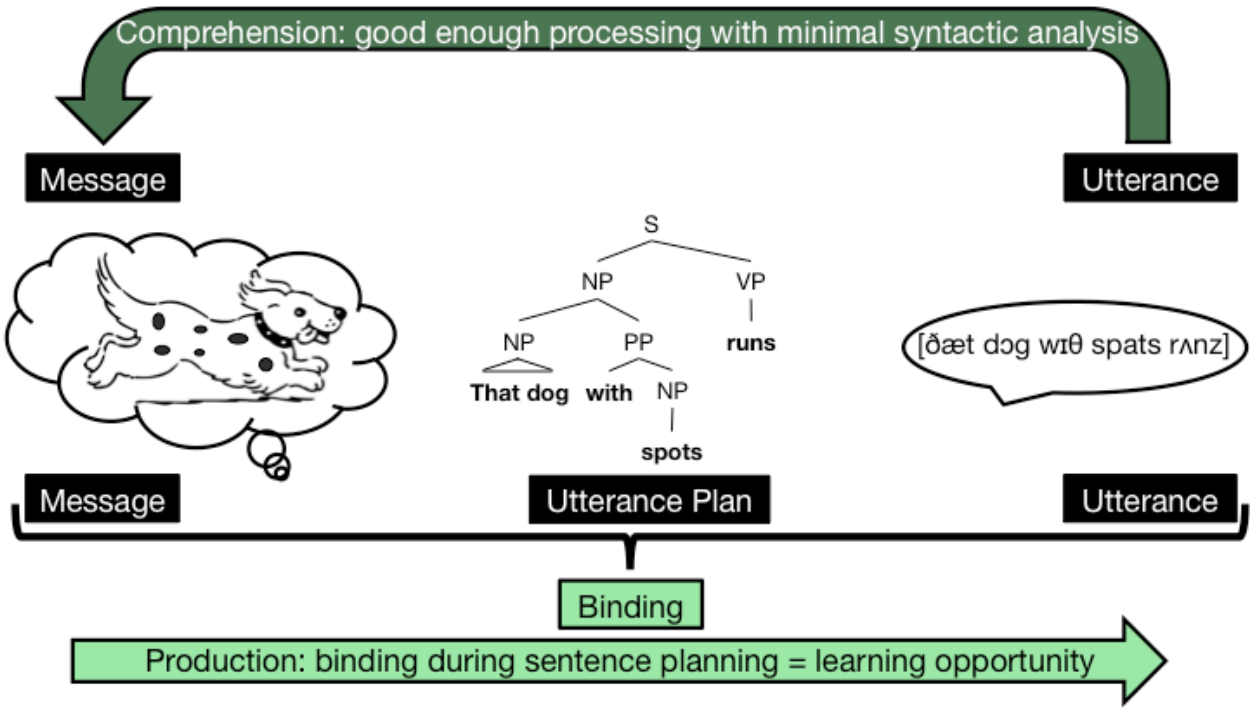


Fig. 1.

Language processing. Language production is the act of turning an idea into a structured utterance, involving generation of structure from long-term knowledge, which relies on recall from memory. Temporary maintenance of word order, bound to the conceptual representation of the message, could be a route for improved learning of multi-word dependencies. During comprehension, perceivers may settle for a "good enough" interpretation without a detailed analysis of all syntactic dependencies, which can get by with mostly recognition-based processing.

Method

Participants

A total of 125 native speakers of English from the University of Wisconsin-Madison received course credit or payment for participating. Based on a power analysis and pilot testing, the goal was 100 participants (50 per learning condition) who scored above threshold on a

vocabulary test. To reach this threshold with leeway to remove non-performing participants, 62 were assigned to the comprehension learning condition and 63 participants were assigned to the production learning condition. Three participants (2 comprehension, 1 production) were unable to finish the experiment.

Materials

Vusu	**Fumusu**	**Teepusu**	**Traw**	**Ot**	**Divusu**	**Kredel**
Determiner	Adjective	Noun	Markings	Preposition	Verb	Location
The-k-pl	Yellow-k-pl	Teep-k-pl	Curves	With	Grow-k-pl	Mountain

The two kind yellow teeps with curved lines grow bigger at the location with the mountain.

Fig. 2.

Artificial world. The first and last frames of a video, the sentence describing it, the grammatical category of each word, the English translation of each word (k-pl = kind-looking, plural) and the sentence in English. Agreeing suffixes are underlined**.** In the experiment, the assignment of words to meanings was different for each participant, and the language was always auditorily presented, never written.

Language and visual word. A cartoon world of monsters situated on alien landscapes was created, including both still pictures and short videos. A language describing the entities, locations, and actions in the world contained 20 root words and four suffixes. All sentences had the structure shown in Fig. 2. See the Supplementary Online Materials – Reviewed (SOM-R) for

more detail on the materials. All materials, including the code to run the experiment in PsychoPy (Peirce, 2007), are available online (https://osf.io/74kqe/#).

Dependencies. The language contained two deterministic agreement dependencies. Suffixes on four word types (determiner, adjective, monster, and verb) varied with the noun number (singular, plural) and monster type (kind-, scary-looking), a semantic category similar to gender or classifier morphology in some natural languages. The *usu* suffix in Fig. 2 indicates a kind looking monster (the *us* part of the suffix) and plural with the final *u* (Table 1).

Table 1

The Four Possible Suffixes and Their Meaning

	Singular	Plural
Kind	-us	-usu
Scary	-ok	-oko

We also introduced a probabilistic dependency, in which monsters tended to be marked with either striped or dotted markings more frequently based on their semantic type. We explored the possibility that production experience would boost such learning, but there was no evidence of any learning of this dependency in our study. While we cannot interpret these null results, it is noteworthy that Amato and MacDonald (2010) found learning of a similar probabilistic dependency with sensitive reading measures but not with measures similar to ones used in the present study. Details about the probabilistic dependency manipulation and results for it can be found in the Supplementary Online Materials – Unreviewed (SOM-U).

Training Procedure

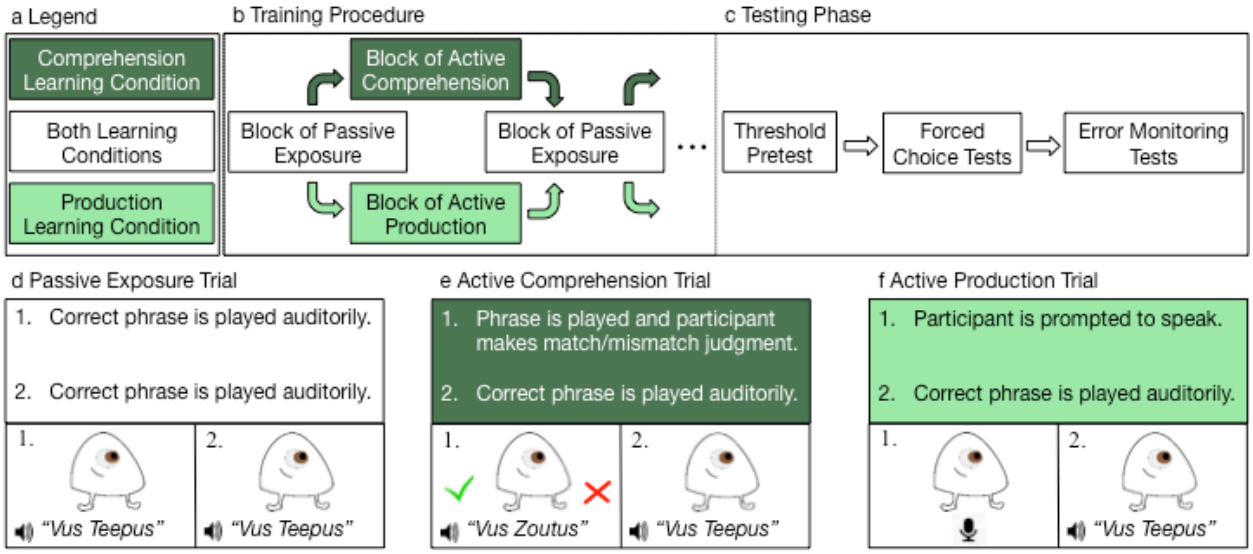


Fig. 3.

Flow of experimental procedure. Training consists of 31 blocks of alternating passive and active learning trials (see Table 1 in SOM-R for more details). After training, participants completed three tests of learning. The two learning conditions were identical in passive exposure blocks and all comprehension tests; the groups' experience differed only in the active learning blocks.

Training (Fig. 3) consisted of blocks of passive exposure trials, interleaved with blocks of either active comprehension trials (comprehension learning condition) or active production trials (production learning condition). All participants received 78 passive exposure trials divided into 14 blocks of 2 to 6 trials each in which a picture or video was paired with two auditory presentations of a word, phrase or sentence in the language that matched the image (Fig. 3d, see Table 1 in SOM-R for more details). For all participants, language training began with a block of still pictures of uncolored unmarked monsters described by single words, and new vocabulary was gradually added in each block until all elements were combined into full sentences, as in Fig. 2. All participants received 96 active learning trials divided into 17 blocks of 2 to 6 trials of the same type each.

Learning conditions. In the *active comprehension* task (Fig. 3e), participants saw a picture and heard a phrase in the novel language, and they indicated with a keypress whether the phrase matched the picture. Half of the trials in each block were mismatches. Feedback on response accuracy was presented onscreen. Regardless of accuracy, feedback was followed by a repetition of the auditory phrase, accompanied by its matching picture.

The *active production* (Fig. 3f) task prompted participants to describe a picture aloud in the artificial language. Responses were recorded. Participants pressed a key after speaking, then heard the phrase that correctly described the picture. The picture remained onscreen throughout the trial, and the correct phrase was presented independent of the accuracy of their production.

Language production and comprehension differ in many dimensions, but our procedure reduced some of these differences. First, amount of listening experience (factor 1, Table 2) was more balanced than in typical comprehension/production comparisons: Comprehension participants heard a phrase that sometimes matched the picture, and production participants heard their own production, which also often was not correct. Furthermore, all participants heard the correct phrase after they made a judgment or said a phrase, providing them with a correct pairing of language and picture. The tasks were also designed to minimize differences in attention and task-relevant choices (factor 2, Table 2), as both tasks required an overt response to a picture. Comprehension trials involved a match/mismatch choice whereas production trials involved more open-ended production choices. Both tasks thus substantially differed from the passive exposure trials, which required no response.

The two learning tasks still capture important inherent differences between production and comprehension. Production involves recall, whereas comprehenders, especially in naturalistic learning settings where a context is provided, can rely on recognition (Fig. 1), with

known consequences for vocabulary learning (Roediger & Karpicke, 2006). In order to investigate the benefits of production beyond vocabulary learning, we controlled for potential differences in vocabulary learning between the two conditions in two ways described in factor 3, Table 2.

Table 2

Three Factors that Differ between Comprehension and Production.

Factor	Explanation	Methods to Reduce Factor in Current Study
1. Double Experience	Every production yields perception experience (e.g., hearing oneself talk).	Comprehension and production conditions both involved hearing a phrase that did/didn't match a picture, then hearing correct matching phrase.
2. Attention & Task-relevant Choices	Production is more attention-demanding than comprehension and inherently involves making choices about what to say.	Active task involving task-relevant choices in both production and comprehension training.
3. Improved vocabulary learning in production	Comprehension requires recognition of the linguistic signal, while production requires recall, which has been shown to improve vocabulary learning. Any investigation of effects on grammar learning should accommodate potential vocabulary learning differences.	Performance threshold for participant inclusion; Vocabulary score as a covariate to test additional benefit of production on grammar learning beyond any benefit for vocabulary learning.

Testing Phase (Fig. 3c)

Threshold pretest. After training was completed, participants were assessed on their learning of the content words of the artificial language, to exclude low-performing participants (factor 3, Table 1). The test consisted of 18 trials in which one word was presented together with two pictures of the same category (e.g., two monsters), testing all content words of the language.

Comprehension participants (M = 16.5, SD = 2.1) and production participants (M = 16.3, SD = 2.2) did not differ in accuracy on this test, $t(120) < 1$. Based on pilot testing, we set a threshold of 15/18 correct (83%) for inclusion in the main analyses. A total of 52 out of 60 comprehension participants and 52 out of 62 production participants met this threshold. All further analyses reported here included data only from these 104 participants scoring above threshold. However, results remained the same when data from the 18 low-scoring participants were included.

Forced choice tests. All participants completed forced choice comprehension trials, similar in format to the comprehension group's active comprehension trials during learning. In each trial, participants saw two pictures on the screen and heard a phrase. They were instructed to choose the picture matching the phrase as quickly as possible, using a keypress, which ended the trial. The dependent variables for these tests were accuracy and reaction time (RT), measured from the onset of the first word in the auditory phrase that could be used to identify the correct picture (words marked with arrows in Fig. 4). Because the participant could respond at any point in the trial, responses occurring before this critical word were recorded as negative RTs. Each trial assessed a particular aspect of language knowledge (vocabulary, suffix meaning, etc.) by virtue of the contrast between the target and the foil picture, and trials were randomly intermixed.

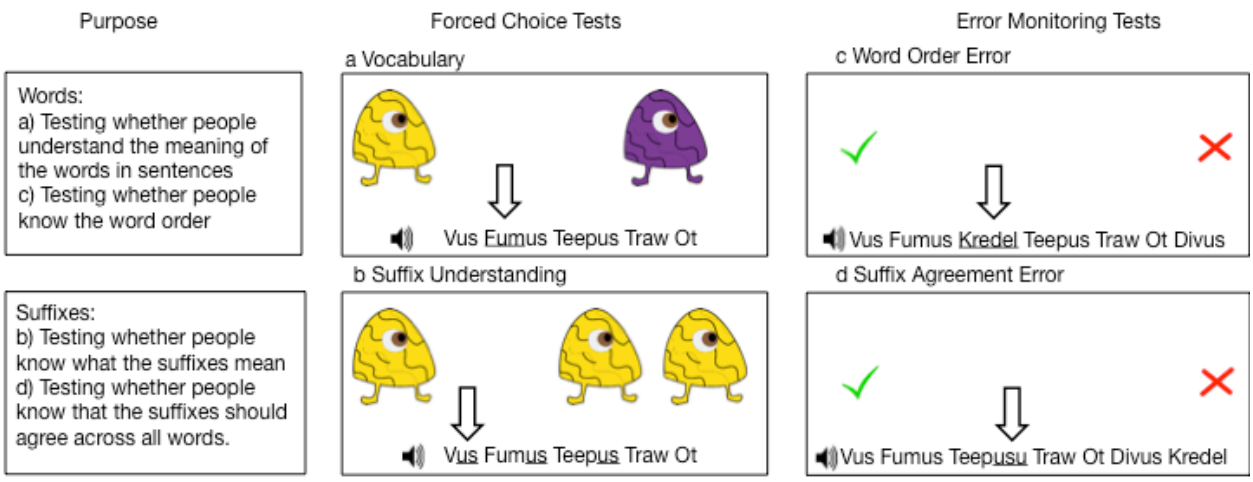


Fig. 4.

Overview of main tests. Participants never saw the language written, they heard only auditory phrases. In the forced choice tests (a,b), participants heard a phrase and chose between the two pictures. In the error monitoring tests (c,d), participants heard a sentence and made a grammaticality judgment. Underlining and arrows indicate the critical word(s) for the participant's response.

Vocabulary test (18 trials). Participants heard a phrase and chose between two pictures that differed in only the meaning of one critical content word. In the example in Fig. 4a, participants heard a five-word phrase and chose between two pictures differing only in color (word 2 of the phrase). As in the threshold pretest, all 18 content words of the language were tested as a trial-critical word. Unlike the pretest, these test items were embedded in full sentences, yielding a difficult auditory sentence comprehension task. We nonetheless expected the groups to perform similarly, as low-performing participants had been excluded by the pretest. Our aim with these trials was both to compare vocabulary and auditory comprehension across groups and also to provide a covariate (vocabulary score) that would allow us to examine learning of grammatical features across groups while controlling for vocabulary learning (factor

3, Table 2). Foils in this test were always within semantic type for monsters and within marking type, so that knowledge of the suffixes and probabilistic monster-marking regularity could not help choose the right picture.

Suffix understanding test (24 trials). Participants heard a phrase and had to choose between two pictures that differed either in monster number (12 items; example in Fig. 4b) or in semantic monster type (12 items). Because the monster word is preceded by two suffixed words (determiner and color) that carry number and semantic information, it is possible to identify the correct picture before hearing the monster word (which also conveys the correct response). The resulting within-subject predictor for number/semantic items did not interact with our main learning condition predictor, and is thus further discussed only in SOM-R.

Error monitoring tests. An error monitoring task assessed participants' sensitivity to violations of language patterns. This test differed from both learning conditions in that there was no picture presented, but the participant's task, judging the correctness of a sentence, was similar to the judgment task in the active comprehension condition. Trials assessing word order and suffix agreement (Fig. 4) were randomly intermixed with 44 grammatically correct sentences. None of these sentences had been presented during training; for sentences with errors, the correct version also had not been presented in training. Participants heard a sentence and pressed a key as quickly as possible to indicate whether the sentence contained an error or was grammatical. The dependent variables were accuracy and reaction time (RT). For each trial, the critical word was defined as the first word that was incorrect (words marked with an arrow in Fig. 4 c-d); in fully correct sentences, the critical word was the last word. Participants occasionally responded before hearing the critical word, leading to some negative RTs.

Word order error test (32 trials). We included trials with four different ungrammatical word orders, each of which had one word in an ungrammatical position (example in Fig. 4c).

Suffix agreement error test (48 trials). Participants heard sentences with different kinds of agreement errors, with one suffix that did not match the other three in the sentence. In the example in Fig. 4d, the mismatching suffix *usu* is plural, whereas the other suffixes are *us*, singular. The within-subject predictors for location of the mismatching suffix never interacted with our main learning condition predictor, and so error type is discussed only in SOM-R.

Predictions. Due to the enhanced serial processing requirements of production, we expected the production group to outperform the comprehension group on tasks with a serial dependency, both word order and suffix agreement across words. If transfer does not occur, the comprehension group should outperform the production group, both because all tests assessed comprehension, and because the testing procedures were more similar to the tasks performed in the comprehension learning group than the production group.

Results

Data Processing

RTs were analyzed with mixed effects regression analyses in R (R Core Team, 2016). Accuracy data were analyzed with mixed effects logistic regression analyses using the lme4 package (Bates, Maechler, Bolker, & Walker, 2015). No trials were removed for the accuracy analysis. In the RT analysis, trials in which the participant responded incorrectly or before the critical word (negative RTs) and RTs more than 3 SD above a participant's own mean were removed, leaving 78% of trials for the RT analysis. Following Barr, Levy, Scheepers and Tily (2013), regression models initially included a random intercept by participants as well as random slopes by participants for all within-subjects predictors (e.g. itemtype). The model for accuracy

in the agreement error test did not converge, so we gradually simplified it (Barr et al., 2013), leading to a model without an intercept but with random slopes by participants; see SOM-R for all statistical models and their outputs. Model predictions for the learning condition predictor for each test are plotted in Fig. 5. They are based on the full model but collapsed over other predictors by taking the average for all within subject predictors (e.g., itemtype), because our main manipulation of learning condition (production versus comprehension) never interacted with any within-subjects predictors; those predictors are not discussed further. All data and analyses are freely available online (osf.io/bbf3c).

Forced Choice Tests

Comprehension participants and production participants did not significantly differ in accuracy on the forced choice vocabulary test. However, there was a range of individual differences in proportion correct (Fig. 5a). Not surprisingly, performance on this task (vocabulary score) was a reliable predictor of accuracy and RT on almost all other tests, indicating that word comprehension in auditory phrases is associated with higher accuracy and shorter RTs on other auditory comprehension tests. Specific results for each test can be found in the results table (Table 2 in SOM-R). Importantly, because we included each participant's score as covariate in all subsequent analyses, all further regression results hold true over and above potential vocabulary score differences between participants.

Production participants were significantly faster than comprehension participants on the vocabulary test items they answered correctly. Production participants were also significantly more accurate and had shorter RTs than comprehension participants on suffix understanding items, which is evidence that production participants had a better understanding of what the suffixes mean.

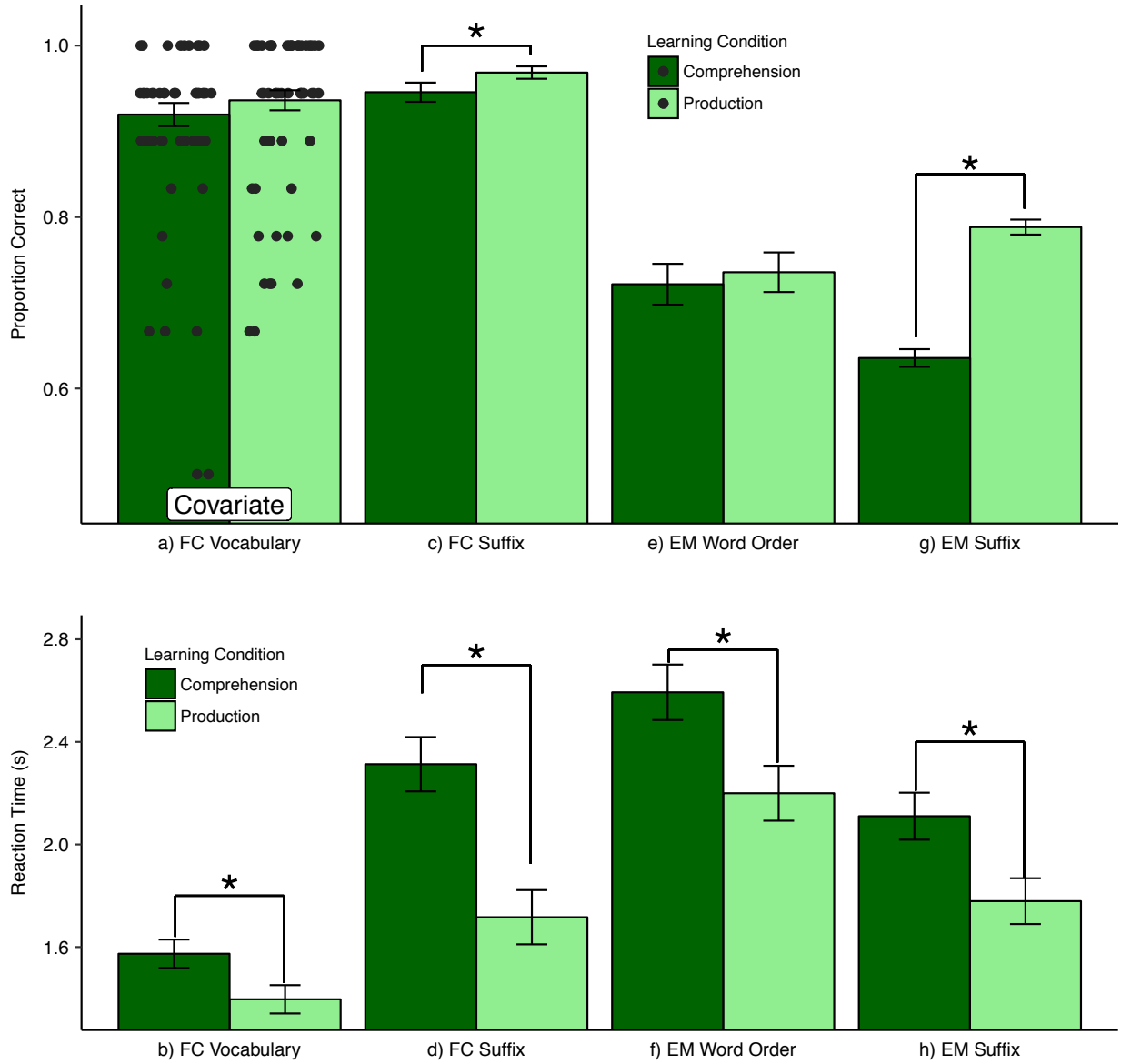


Fig 5.

Overview of comprehension tests results. (FC = Forced Choice, EM = Error Monitoring, * p<.05). Bars show model predictions, error bars show 95% CI. Accompanying table with regression models for all tests can be found in the supplementary materials. In a) the dots represent proportion correct for individual participants, which is used as a covariate in all other regression analyses.

Error Monitoring Tests

We calculated a d-prime score for each participant, reflecting their sensitivity to grammar (discriminating between correct vs. incorrect word order and agreement), and compared d-prime scores between learning conditions. Production participants (M = 2.4, SD = 1.1) were overall significantly more sensitive than comprehension participants (M = 1.8, SD = 1.1) across the two error types $t(102) = 2.36$, $p = 0.020$.

Separate word order and suffix agreement tests yielded similar results, with production outperforming comprehension participants in both accuracy and speed, with the exception of no reliable differences in accuracy for the word order test.

Discussion

Production-focused training yielded superior learning and comprehension of a novel language, across a variety of language features and task demands, compared to training focusing on comprehension itself. Importantly, production's learning advantage went beyond the word level: even after controlling for vocabulary knowledge, production participants were both faster and more accurate on tests of grammar comprehension.

The balancing of production and comprehension conditions allows us to take steps in identifying possible mechanisms underlying production's beneficial effects. While lexical retrieval (the recall of words from long-term memory) is likely to be a powerful component of production's learning benefits, other aspects of utterance planning may also be important contributors. Language production begins with a conceptual message that the producer aims to communicate. This message, fully known to the producer, is activated throughout utterance planning and execution, promoting binding over all parts of the utterance (Savill et al., 2017) (Fig. 1). This situation should afford a stronger learning opportunity than in comprehension, where the input signal and the message that the comprehender gleans from it unfold over time.

Language production also requires the generation of an utterance plan, and because planning precedes execution by some time, planning entails maintaining information in working memory (Brown-Schmidt & Konopka, 2015). Indeed, MacDonald (2016) argued that the utterance plan *is* the maintenance portion of verbal working memory. The temporary maintenance, serial ordering, and binding across the different linguistic levels that occurs during utterance planning provides benefits for learning inter-word grammatical, conceptual, and phonological relationships. These relationships may underlie our finding that production benefits grammatical learning beyond vocabulary knowledge. Our control for vocabulary knowledge in grammar learning is a first step to exploring the different kinds of learning opportunities that production processes afford.

Our results constrain theoretical positions on verbal memory and learning in several ways. First, they show that the benefit of production on language learning need not depend on an additional potential learning experience in the form of hearing oneself speak (see MacLeod & Bodner, 2017, for other comprehension-production differences in word lists). The current study balanced listening experience across learning conditions and still found benefits for production over comprehension learning tasks. Second, our results expand the reach of the testing effect (Roediger & Karpicke, 2006): we posit that language production inherently has important learning benefits that have been associated with testing. Full language production involves recall of words from long-term memory and assembly of sentence structure, whereas comprehension relies more heavily on recognition. A more limited production task, repetition of another's utterance, does not require full generation of language from memory and appears to have reduced learning benefits in vocabulary learning compared to full, generative production (Kang, Gollan & Pashler, 2013; Middleton, Schwartz, Rawson & Garvey, 2015). Third, we extend for

the first time a production-based testing effect beyond single words, and show that language production is superior to comprehension training in learning and comprehension of grammar, even after imposing controls for differences in word knowledge. Consistent with our result with spoken language, there is evidence that retrieval practice improves conceptual learning from texts (Karpicke & Blunt, 2011), also suggesting an important role for retrieval/production practice in relational learning, whether it is making inferences about concepts or learning grammatical regularities. Fourth, our results corroborate findings that spelling practice on difficult written words improves reading speed on these words (Ouellette, Martin-Chang & Rossi, 2017); though not explicitly phrased as such, these findings provide another example of production practice improving comprehension in a different but related modality.

The finding that production training improves subsequent comprehension performance more than comprehension practice itself provides a clear example of learning transfer, where experience with one task yields subsequent benefits on a different task (Fan, Turk-Browne, & Taylor, 2016; Green et al., 2015). This transfer effect is most readily understood as reflecting shared representations between comprehension and production. Future work should examine the extent to which benefits for production extend to other levels of language perception and comprehension beyond the lexical and grammatical aspects studied here, because evidence is mixed concerning benefits and costs to production at the level of speech sound perception (Baese-Berk & Samuel, 2016; Bixby, 2017; Leach & Samuel, 2007).

Our findings also have implications for language instruction, including Krashen's (2003) input hypothesis, which holds that language learning is driven by comprehension practice, not production. Studies of classroom language learning have supported this claim, showing that comprehension practice leads to better production performance, but not vice versa (VanPatten &

Cadierno, 1993). These results initially seem to directly contradict our own, but a key difference is in how "production" is instantiated. Whereas Krashen and colleagues associate "production" with repetition of teacher input and spoken grammar drills, the production learning in our experiment involved generation of meaningful language, and we showed that such practice is beneficial. Because the mantra that "comprehension is better than production practice" is widespread in some approaches to second language instruction (Krashen, 2003), it will be important to distinguish repetition from more generative production in both future research and recommendations to instructors.

This work may also illuminate effects of child production and comprehension in first language acquisition. Children from economically disadvantaged households tend to have reduced language experience compared to those in more affluent households, with consequences for vocabulary development and educational outcomes (Hart & Risley, 1995). While differences are commonly framed in terms of comprehension - the "thirty million word gap" in the amount of language the child hears, other studies suggest a key role for the child's own production experience. Zimmerman et al. (2009) found that the number of turns in adult-child conversations was a better predictor of language development than language input (comprehension experience). In conversational exchanges, the child not only hears adult input but also produces language. Beyond other stimulating and engaging aspects of conversational turns, the present results suggest that affording the child opportunities to produce language may provide the learning benefits inherent in language production.

Author Contributions

E. W. M. Hopman and M. C. MacDonald designed the study. E. W. M. Hopman conducted the study and analyzed the data, under supervision of M. C. MacDonald. E. W. M. Hopman and M. C. MacDonald interpreted the findings and wrote and revised the manuscript.

References

Amato, M.S., & MacDonald, M. C. (2010) Sentence processing in an artificial language: Learning and using combinatorial constraints. *Cognition, 116*(1), 143-148.

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.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-279.

Bates, D., Maechler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67(1),* 1-48.

Bixby, K. N. (2017). *Production effects on perception: How learning to produce sound changes auditory perception* (Unpublished doctoral dissertation). University of Rochester, Rochester, New York.

Bock, K., Dell, G. S., Chang, F., & Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, *104*, 437–458.

Boiteau, T.W., Malone, P.S., Peters, S.A., & Almor, A. (2014). Interference between conversation and a concurrent visuomotor task. *Journal of Experimental Psychology: General*, *143*, 295–311.

Branigan, H. P., Pickering, M. J., & McLean, J. F. (2005). Priming prepositional-phrase attachment during comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 468-481.

Brown-Schmidt, S., & Konopka, A. E. (2015). Processes of incremental message planning during conversation. *Psychonomic Bulletin & Review, 22*(3), 833-843.

Carter, M. J., & Ste-Marie, D. M. (2017). Not all choices are created equal: Task-relevant choices enhance motor learning compared to task-irrelevant choices. *Psychonomic Bulletin & Review*, *24*(6), 1879–1888.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294.

Fan, J. E., & Turk-Browne, N. B. (2013). Internal attention to features in visual short-term memory guides object learning. *Cognition, 129,* 292-308.

Fan, J. E., Turk-Browne, N. B., & Taylor, J.A. (2016). Error-driven learning in statistical summary perception. *Journal of Experimenal Psychology: Human Perception and Performance, 42(2),* 266.

Ferreira, F., & Patson, N. D. (2007). The 'good enough' approach to language comprehension. *Language and Linguistics Compass, 1*, 71-83.

Green, C. S., Kattner, F., Siegel, M. H., Kersten, D., & Schrater, P. R. (2015). Differences in perceptual learning transfer as a function of training task. *Journal of Vision, 15,* 5.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences, 23,* 1-71.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children.* Paul H. Brookes Publishing.

Kang, S.H.K., Gollan, T.H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin Review, 20,* 1259-1265

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science, 319,* 966-968.

Karpicke, J. D. & Roediger, H.L. (2008). The critical importance of retrieval for learning. *Science, 331,* 772-775.

Krashen, S. D. (2003). *Explorations in language acquisition and use* (pp. 1-27). Portsmouth, NH: Heinemann.

Leach, L., & Samuel, A. G. (2007). Lexical configuration and lexical engagement: When adults learn new words. *Cognitive Psychology, 55,* 306-355.

MacDonald, M. C. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science*, *25*, 47-53.

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26,* 390-395.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. *Journal of Experimental Psychology: General*, *144*(2), 447.

Middleton, E. L., Schwartz, M. F., Rawson, K. A., & Garvey, K. (2015). Test-enhanced learning versus errorless learning in aphasia rehabilitation: testing competing psychological principles. *Journal of Experimental Psychology: Learning, Memory and Cognition, 4,* 1253-1261.

Ouellette, G., Martin-Chang, S., & Rossi, M. (2017). Learning from our mistakes: Improvements in spelling lead to gains in reading speed. *Scientific Studies of Reading, 21*(4), 350-357.

Peirce, J. W. (2007). PsychoPy – psychophysics software in Python. Journal of Neuroscience Methods, 162(1), 8-13.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*, 329–347.

R Core Team (2016). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2014.

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning. *Psychological Science, 17*, 249-255.

Savill, N., Ellis, R., Brooke, E., Koa, T., Ferguson, S., Rojas-Rodriguez, E., Arnold, D., Smallwood, J., & Jefferies, E. (2017). Keeping it together: Semantic coherence stabilizes phonological sequences in short-term memory. *Memory & Cognition*, 1–12. https://doi.org/10.3758/s13421-017-0775-3

Segaert, K., Menenti, L., Weber, K., Petersson, K. M., & Hagoort, P. (2011). Shared syntax in language production and language comprehension—an fMRI study. *Cerebral Cortex*, *22*(7), 1662-1670.

VanPatten, B. (2013). Input Processing. In S.M. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition* (pp. 268-281). Routledge.

VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition*, *15*(2), 225-243.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 135-144.

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., & Yapanel, U. (2009). Teaching by listening: The importance of adult-child conversations to language development. *Pediatrics, 124,* 342-349.

Supplementary Online Materials - Reviewed
Method

Materials

Visual world. There were three each of two types of monsters, which we term kind and scary looking (Fig. 1, nouns). Kind looking monsters (Fig. 1, noun column, top three rows) have rounded shapes, one brown eye, are smiling, and have two feet. Scary looking monsters (Fig. 1, noun column, bottom three rows) have angular shapes, five green eyes, a fanged mouth, antenna and six legs. All monsters appeared approximately equally in two different colors (Fig. 1, adjectives). Additionally, each monster could have two types of striped markings (curves and lines) and both smaller and larger spots (Fig. 1, markings). The monsters performed three actions: growing, moving up, or moving left to right (Fig. 1, verbs). There were three possible locations: mountains, craters and river (Fig. 1, locations).

Determiner	Adjective		Noun		Markings		Preposition	Verb-		Location	
V-	Fum-		Teep-		Traw		Ot	Div-		Kredel	
	Saf-		Zout-		Plim			Pav-		Chaftem	
			Weem-		Chag			Zev-		Hullem	
			Mog-		Stam						
			Ket-								
			Pex-								

Figure 1. Overview of all words in the artificial language by word type and the visual elements they refer to. A dash at the end of a word indicates that the word gets a suffix.

Artificial language. Fig. 1 shows an overview of all the words in the artificial language. A native speaker of English recorded all of these words (with all possible suffixes, where relevant) in a soundproof booth. Words were spoken individually with neutral intonation and English pronunciation. Longer phrases and sentences were created by concatenating recordings of individual words. The average length of a 7-word-sentence was 5338 ms.

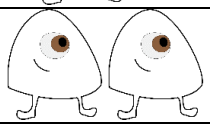
Counterbalancing. Mapping of words to visual referents were randomized *within word type* for each participant, as was the assignment of suffix (us/usu or ok/oko) to semantic monster type for the suffix agreement dependency and the assignment of marking type to monster type for the probabilistic dependency. The mappings used in this write-up (Fig. 1) are just an example of a possible assignment. During exposure and test, all stimuli are balanced within type; e.g. both colors are seen equally often. Furthermore, first order contingencies are minimalized so that the only regularity is the probabilistic monsters-markings co-occurrence; e.g. each monster appears equally often in each color.

Training Procedure

Table 1

Overview of Training Procedure

[a]Pilot training showed that when 6 new vocabulary words were introduced at the same time (as happened with the monsters and the verbs + landscapes), this was hard to learn all at once, so an extra active learning block was added to help participants learn the vocabulary in these situations.

Nr of trials	Stimulus type	Block type	Example utterance	Example visual stimulus	Rationale
6	singular monster vocabulary	passive exposure	*"Vus Teepus"*		There are 6 different monsters.
6	singular monster vocabulary	active learning			
6	singular monster vocabulary	active learning[a]			
6	plural monster vocabulary	passive exposure	*"Vusu Teepusu"*		
6	plural monster vocabulary	active learning			
6	plural monster vocabulary	active learning[a]			
2	color vocabulary	passive exposure	*"Fum"*		There are 2 different colors.
2	color vocabulary	active learning			
6	colored monster	passive exposure	*"Vus Fumus Teepus"*		One block of 6 trials is enough to balance the 6 monsters and 2 colors.
6	colored monster	active learning			
4	markings vocabulary	passive exposure	*"Traw"*		There are 4 different markingss.
4	markings vocabulary	active learning			
6	colored monster with markings	passive exposure	*"Vus Fumus Teepus Traw Ot"*		Two blocks of 6 trials are necessary to balance the 6 monsters, 2 colors and 4 markings.
6	colored monster with markings	active learning			
6	colored monster with markings	passive exposure			
6	colored monster with markings	active learning			
3+ 3	verb and landscape vocabulary	passive exposure	*"Kredel"*		There are 3 different verbs and 3 different landscapes.
3+ 3	verb and landscape vocabulary	active learning			
3+ 3	verb and landscape vocabulary	active learning[a]			
6	full sentence	passive exposure	"Vus Fumus Teepus Traw Ot Divus Kredel"	  (first and last frame of video are shown here)	This is the main part of training. Based on pilot testing, 6 blocks each of passive exposure and active learning was enough for participants to learn the grammar without getting to ceiling.
6	full sentence	active learning			
6	full sentence	passive exposure			
6	full sentence	active learning			
6	full sentence	passive exposure			
6	full sentence	active learning			
6	full sentence	passive exposure			
6	full sentence	active learning			
6	full sentence	passive exposure			
6	full sentence	active learning			
6	full sentence	passive exposure			
6	full sentence	active learning			

Testing After Learning

Suffix agreement error types. Suffix agreement errors were created by either mismatching the suffix number or the suffix semantic type. Adjacent errors were created by putting the mismatching suffix on the monster word (the third word of the sentence, as in the

example), next to the determiner and color word that also both get a suffix. Non-adjacent errors were created by putting the mismatching suffix on the verb, the fifth word of the sentence.

Results

Table 2 shows the results of the regression analyses for each test. RT data are analyzed with mixed effects regression models and accuracy data with logistic models. In all analyses, learning condition was coded with comprehension as -0.5 and production as 0.5. Thus, a positive significant coefficient for learning condition means that participants in the production learning condition were more accurate than participants in the comprehension learning condition. A negative coefficient for learning condition in RT analyses means that participants in the production learning condition were faster (had a lower RT) than participants in the comprehension learning condition. All results for the learning condition predictor are also depicted visually in Fig. 5 in the main text. The vocabulary score predictor, an individual's proportion correct on the vocabulary understanding in phrases test, ranges from 0.4 to 1 in our sample. All other predictors are within-subjects, reflecting different types of items within a given test. Their coding is explained in table notes. None of these within-subjects predictors ever interacted with learning condition, and the results are thus of less interest here and not interpreted or discussed further.

Table 2

Mixed effects (logistic) regression model results for all analyses depicted in Fig. 5 in the main text.

		Accuracy				RT				
Test	Fixed effects	Coefficient	Standard Error	*z* value	*p* value	Coefficient	Standard Error	*F*	Error df	*p* value
Forced Choice Vocabulary	Intercept	2.56	0.14	18.18	< .001	1.49	0.04	1443.17	102.0	< .001
	VocabularyScore[a]	-	-	-	-	-1.08	0.40	7.51	124.2	0.007
	LearningCondition	0.25	0.25	0.99	> .250	-0.18	0.08	5.17	100.3	0.025
Forced Choice Suffix Understanding	Intercept	3.14	0.18	17.34	< .001	2.01	0.07	727.21	102.1	< .001
	VocabularyScore	4.49	1.19	3.78	< .001	-0.86	0.69	1.55	109.3	0.214
	LearningCondition	0.57	0.28	2.07	0.039	-0.60	0.15	15.83	102.2	< .001
	Itemtype[b]	1.0	0.33	3.02	0.003	0.09	0.08	1.31	109.9	>.250
	LearningCondition*Itemtype	-0.47	0.49	-0.97	>.250	-0.21	0.16	1.88	109.9	0.173
Error Monitoring Word Order	Intercept	0.99	0.08	11.74	< .001	2.40	0.08	991.27	100.3	< .001
	VocabularyScore	3.22	0.77	4.20	< .001	-1.10	0.78	2.04	107.0	0.160
	LearningCondition	0.07	0.17	0.42	>.250	-0.40	0.15	6.70	99.6	0.011
Error Monitoring Suffix Agreement[d]	Intercept	0.94	0.03	27.18	< .001	1.94	0.06	922.85	102.7	< .001
	VocabularyScore	3.65	0.30	12.17	< .001	-1.05	0.64	2.70	110.2	0.104
	LearningCondition	0.76	0.07	11.07	< .001	-0.33	0.13	6.70	102.2	0.011
	Adjacency[c]	-1.08	0.07	-15.70	< .001	0.07	0.05	1.96	100.8	0.165
	Itemtype[b]	0.32	0.07	4.63	< .001	-0.24	0.03	51.49	99.9	< .001
	LearningCondition*Adjacency	0.06	0.14	0.40	> .250	-0.08	0.10	0.72	100.8	> .250
	LearningCondition*Itemtype	-0.08	0.14	-0.54	> .250	0.05	0.07	0.49	99.9	> .250
	Adjacency*Itemtype	-0.74	0.14	-5.40	< .001	0.24	0.06	13.68	95.8	< .001
	3-way Interaction	0.20	0.27	0.74	> .250	-0.09	0.13	0.50	95.7	> .250

Note. Accuracy was analyzed with mixed effects logistic regression models (glmer) in R, RT with mixed effects regression models (lmer) in R. Unless otherwise indicated, all models included the full random effects structure.

[a]Since VocabularyScore was based on the accuracy data in the Forced Choice Vocabulary Test, it was not a predictor for the accuracy model itself.

[b]Itemtype: the agreement error test included both semantic (0.5) and number (-0.5) agreement errors.

[c]Adjacency: the agreement error test included both non-adjacent (0.5) and adjacent (-0.5) agreement errors.

[d] This model did not initially converge. Suggestions from Barr, Levy, Scheepers, & Tily, (2013) were followed, leading to a model without by participant random intercepts, with the following specification: Correct ~ VocabularyScore + LearningCondition*Adjacency*Itemtype + (0 + Adjacency:Itemtype | Participant).

Supplementary Online Materials - Unreviewed
Probabilistic Co-occurrence Depencency
Method

Exposure

Exposure during both training and test (Fig. 1) was set up so that a given type of monsters (e.g. kind looking) occurred more often with a given type of markings (e.g. striped). Assignment of monster and marking types for this regularity was counterbalanced, but for the examples in this write-up we'll have kind looking monsters occur more often with striped markings (83%) than with dotted markings (17%) whereas scary looking monsters occur more often with dotted markings (83%) than with striped markings (17%). This probabilistic dependency is present in both the visual world and in the language. Participants see kind monsters five times more often with striped than with dotted markings. Analogously, participants hear the words for the kind monsters followed more often by the words for striped markings: the transition probability of 'Teepus Traw', a kind striped monster, is five times as high as that for 'Teepus Chag', a dotted version of the same monster. All tests were also set up so that overall, this probabilistic co-occurrence regularity was approximately preserved.

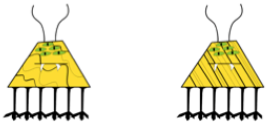
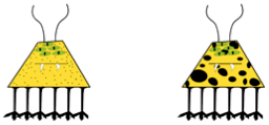
	Stripes *Traw, Plim*	Dots *Chag, Stam*
Nice	84%  *Teepus Traw* *Teepus Plim*	16%  *Teepus Chag* *Teepus Stam*
Scary	16%  *Ketok Traw* *Ketok Plim*	84%  *Ketok Chag* *Ketok Stam*

Fig. 1. The probabilistic co-occurrence dependency implemented in exposure.

Forced Choice Test for sensitivity to probabilistic co-occurrence (24 trials)

To assess whether participants were sensitive to the probabilistic co-occurrences in scenes and words describing them, participants heard a phrase and had to choose between two pictures of the same monster with probable or improbable markings. In Probable trials (12 items), the phrase described the monster with probable markings, whereas in Improbable trials (12 items) the phrase described the monster with improbable markings, leading to 22% improbable trials in the Forced Choice Task.

In order to get these sensitivity to probabilistic co-occurrence items correct, it is enough to understand the meaning of the markings words. However, participants have heard and seen probable combinations of monsters and markings five times more often than improbable combinations, both during training and in test. If they are sensitive to this probabilistic co-occurrence, we would expect them to be relatively slow/inaccurate on the Improbable trials and

relatively fast/accurate on the Probable trials, which would lead to a significant effect of Itemtype. If people are not sensitive to this probabilistic co-occurrence, and only get these items correct based on understanding the meaning of the markings words, we would expect no difference between the two types of items. We expected Production participants to be more sensitive to the probabilistic co-occurrence dependency than Comprehension participants. Thus, we predicted an interaction between Learning Condition and Itemtype.

Error Monitoring Test for sensitivity to probabilistic co-occurrence (44 trials)

Participants heard Probable (24) and Improbable (20) sentences, all of which were grammatical. Probable sentences had monster-markings combinations that had occurred frequently in training, whereas Improbable sentences had low frequency combinations. All other items in the Error Monitoring task (32 Word Order Error and 48 Suffix Agreement Error items) consisted of probable monster-markings combinations, leading to 16% improbable trials in the Error Monitoring Task. With the same rationale as in the Probabilistic Co-occurrence items in the Forced Choice task, we expected an interaction between Itemtype and Learning Condition to show that Production participants were more sensitive than Comprehension participants to the probabilistic co-occurrence.

Results

In the Forced Choice task (Fig. 2a,b) we found no effect of Learning Condition, Itemtype or their interaction in the accuracy data (Table 1). In the RT data we only find a significant effect of Learning Condition, meaning that Production participants are generally faster at all of these items than Comprehension participants (Table 2), but no interaction with Itemtype. Thus, while production participants are generally faster in this test, there was no evidence that either group was sensitive to the probabilistic co-occurrence dependency.

The effects of probabilistic co-occurrences on error monitoring performance (Fi. 2c,d) showed similar results: there were no differences in accuracy, and Production participants were overall faster responders (Tables 3 and 4), but there was no interaction with Itemtype. Again, there was no evidence that participants were sensitive to the probabilistic co-occurrence dependency.

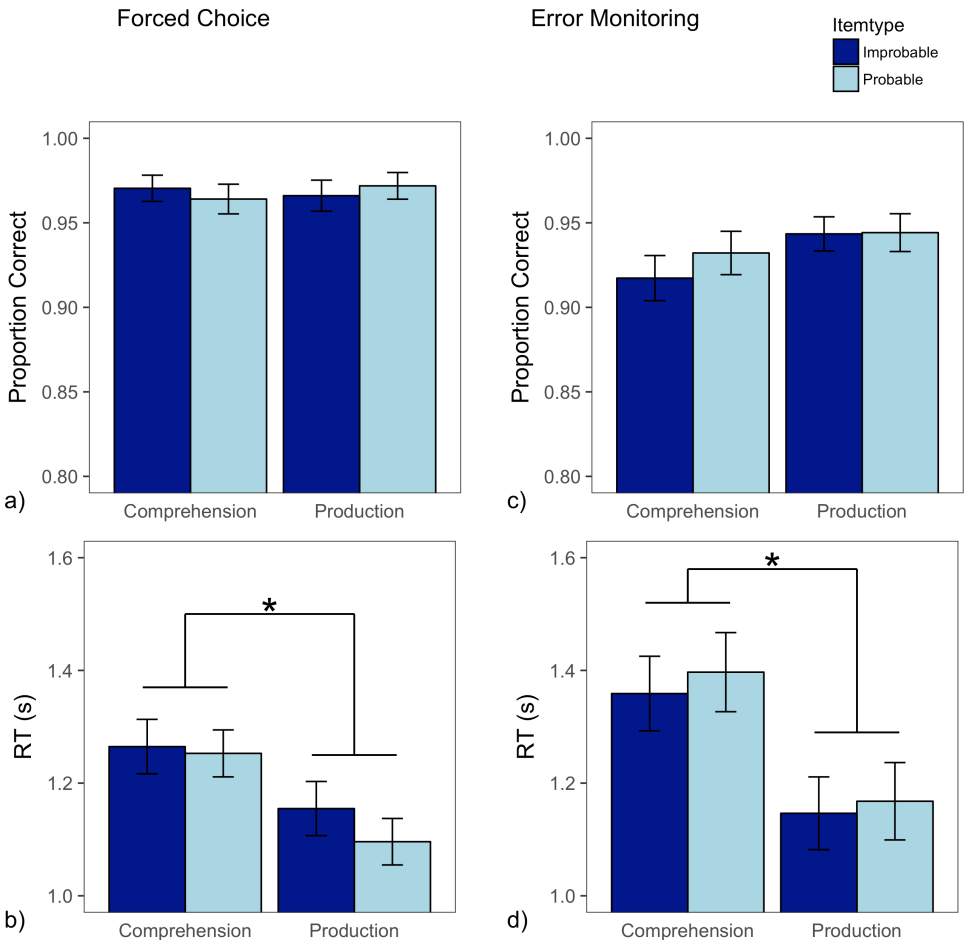


Fig. 2. Accuracy and RT results for the probabilistic co-occurrence test items in the forced choice and error monitoring tasks. Bars show model predictions, error bars show 95% CI, significance of the Learning Condition predictor is indicated with * p<0.05.

Table 1

Accuracy Analysis of the 24 Sensitivity to Probabilistic Co-Occurrence Iems in the Forced Choice Task.

Correct ~ VocabularyScore + LearningCondition*Itemtype + (1 + Itemtype | Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	3.42	0.17	19.60	< .001
VocabularyScore	7.62	1.12	6.78	< .001
Condition	0.06	0.28	0.20	> .250
Itemtype[a]	-0.00	0.25	-0.01	> .250
Interaction	0.40	0.37	1.07	> .250

[a]Itemtype: the suffix understanding test included both probable (+0.5) and improbable (-0.5) items.

Table 2

RT analysis of the 24 Sensitivity to Probabilistic Co-occurrence items in the Forced Choice task.

RT ~ VocabularyScore + LearningCondition*Itemtype + (1 + Itemtype | Participant)

	Coefficient	Standard Error	*F*	Error *df*	*p* value
Intercept	1.21	0.03	1689.72	101.4	< .001
VocabularyScore	-1.34	0.28	22.58	123.4	< .001
LearningCondition	-0.13	0.06	5.15	101.0	0.025
Itemtype[a]	-0.04	0.03	2.10	109.4	0.150
Interaction	-0.05	0.05	0.92	109.3	> .250

[a]Itemtype: the suffix understanding test included both probable (+0.5) and improbable (-0.5) items.

Table 3

Accuracy Analysis of the 44 Sensitivity to Probabilistic Co-Occurrence Items in the Error Monitoring Task.

Correct ~ VocabularyScore + LearningCondition*Itemtype + (1 + Itemtype | Participant)

	Coefficient	Standard Error	*z* value	*p* value
Intercept	2.67	0.13	20.89	< .001
VocabularyScore	6.77	1.03	6.57	< .001
LearningCondition	0.31	0.24	1.27	0.204
Itemtype[a]	0.11	0.13	0.86	> .250
Interaction	-0.20	0.21	-0.95	> .250

[a]Itemtype: the error monitoring task included both probable (+0.5) and improbable (-0.5) items.

Table 4

RT Analysis of the 44 Sensitivity to Probabilistic Co-Occurrence Items in the Error Monitoring Task.

RT ~ VocabularyScore + LearningCondition*Itemtype + (1 + Itemtype | Participant)

	Coefficient	Standard Error	*F*	Error df	*p* value
Intercept	1.26	0.05	752.37	99.8	> .250
VocabularyScore	0.18	0.48	0.15	108.7	_> .250
LearningCondition	-0.22	0.09	5.78	99.2	_0.018
Itemtype[a]	0.03	0.03	1.46	98.6	_0.230
Interaction	-0.02	0.05	0.12	98.5	> .250

[a]Itemtype: the error monitoring task included both probable (+0.5) and improbable (-0.5) items.