

# Journal of Visualized Experiments

## A balanced test of comprehension versus production practice using artificial language learning --Manuscript Draft--

Article Type:	Invited Methods Article - JoVE Produced Video
Manuscript Number:	JoVE59946R2
Full Title:	A balanced test of comprehension versus production practice using artificial language learning
Section/Category:	JoVE Behavior
Keywords:	Behavioral science; experimental psychology; psycholinguistics; Learning; language production; language learning; language comprehension; artificial language learning; learning transfer; open data; open materials
Corresponding Author:	Elise Hopman UNITED STATES
Corresponding Author's Institution:	
Corresponding Author E-Mail:	hopman@wisc.edu
Order of Authors:	Elise W.M. Hopman Mackenzie Ludin Maryellen C. MacDonald
Additional Information:	
Question	Response
Please indicate whether this article will be Standard Access or Open Access.	Open Access (US\$4,200)
Please indicate the <b>city, state/province, and country</b> where this article will be <b>filmed</b> . Please do not use abbreviations.	Madison, Wisconsin, United States of America

**TITLE:**

**A Balanced Test of Comprehension Versus Production Practice Using Artificial Language Learning**

**AUTHORS AND AFFILIATIONS:**

Elise W. M. Hopman, Mackenzie Ludin, Maryellen C. MacDonald

Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

**Corresponding Author:**

Elise W.M. Hopman (hopman@wisc.edu)

**Email Addresses of Co-authors:**

Mackenzie Ludin (mludin@wisc.edu)

Maryellen C. MacDonald (mcmacdonald@wisc.edu)

**KEYWORDS:**

behavioral science, experimental psychology, psycholinguistics, language production, language learning, language comprehension, artificial language learning, learning transfer, open data, open materials

**SUMMARY:**

The goal of this protocol is to evaluate comprehension versus production language learning training through a computer-based experiment in a way that balances inherent task differences between comprehension and production.

**ABSTRACT:**

Research and theories in the field of second language acquisition have long held that language comprehension is a stronger learning experience than language production, especially for learning the grammar of a second language. In contrast, psychology research shows that, at least at the single word level, the opposite is true: language production training, due to the memory processes involved, leads to better learning of words in a second language. The inherent differences between language production and comprehension were not well-balanced in prior research, potentially leading to these conflicting results. Thus, the present study's protocol includes language comprehension and production training tasks that are balanced for listening experience, task-relevant choices, and attention. In the active production task, participants see a picture and are asked to describe it out loud. In the active comprehension task, participants see a picture and hear a phrase. They make a match/mismatch judgment on whether the phrase describes the picture or not. In both conditions, participants hear the correct description of the picture after the task. This full protocol includes computer-based language training in which participants gradually learn an artificial language, building up from single words to full sentences. Training alternates the active production or comprehension tasks with passive exposure to familiarize participants with the language. After training, participants' learning is assessed using several tests that tap into both vocabulary and grammar learning. Versions of this protocol have

been used for learning both artificial and natural languages and have consistently shown that participants in the production condition learn the language better than participants in the comprehension condition. Extensions of this protocol could be used for comparing the effects of comprehension versus production training on different language phenomena in different languages of interest.

## **INTRODUCTION:**

Language learning inevitably involves practice in both comprehension (i.e., listening) and production (i.e., speaking), skills which require different amounts of attention and rely on different memory processes. Focusing on either comprehension or production practice may yield different results in language learning because of the differences in task demands. Research in the field of second language acquisition strongly suggests that to learn the grammar of a second language, comprehension practice is more useful than production practice<sup>1,2</sup>. However, memory research suggests that production practice can boost learning compared to comprehension practice, at least at the single word level. Production practice provides an additional presentation of the material, because the speaker can hear their own pronunciation<sup>3</sup>. Because production practice demands more attention, it leads to greater depth of processing<sup>4</sup>. Production involves making task-relevant choices, because a speaker must choose what to say, and making task-relevant choices increases learning<sup>5</sup>. Finally, comprehension only involves recognition, whereas production relies on recall, which is a stronger learning experience<sup>6</sup>.

Thus, the literature paints a mixed picture of the merits of comprehension versus production training<sup>7</sup>. This may be due to two gaps in the literature that are addressed in the method presented here. First, whereas second language acquisition literature has focused on grammar learning, psychology literature has focused on single word learning. Second, most previous studies did not seek to balance production and comprehension training: while production and comprehension are innately different, some of the task demands can be balanced in order to focus on inherent processing differences.

The method presented here reduces some of the differences between production and comprehension demands to allow for a more direct comparison. The core of the method is to have participants in the comprehension and production conditions practice with the same language materials in different ways. An active training trial gives participants in both conditions practice with the same target image and the target phrase describing it. Each active training trial consist of three parts: the first is a prompt that contains either the target phrase or the target image; the second is the active part, in which the participant responds to the prompt; the third presents the participant with the correct pairing of the target image and the target phrase describing it. The third part with the correct pairing is identical for participants in both conditions. Production participants start with the target image and are asked to describe it out loud in the language they are learning. Comprehension participants start with the target phrase and see an image on the screen that may or may not match the target phrase. Their active task is to make a match/mismatch judgment. Both sets of participants are passively exposed to similar language material before starting the active task, so that they have the same input to draw on for their active task. For both sets of participants, the active response to the prompt (speaking or making

a match-mismatch judgment) is followed by a correct pairing of the image and the phrase that describes it.

In the two tasks contrasted here, the amount of listening experience is more balanced than in previous comparisons between comprehension and production training: comprehension participants hear the target phrase, which may or may not correctly describe the image shown on that trial, while production participants hear their own speech, which also may or may not correctly describe the target image. Both sets of participants then hear and see the target phrase and target image in the final part of the trial to ensure they can learn the language properly. Attention demands are also balanced between the two conditions: both tasks require an overt response to a picture.

This careful balancing of listening experience, task demands, and attention is not generally done in second language acquisition studies comparing production and comprehension training. The comprehension tasks used in that literature<sup>7</sup> are similar to the comprehension training task and the comprehension tests employed after learning: error indication, multiple choice, and matching a picture with input. The crucial difference between this method and most second language acquisition studies comparing production and comprehension training lies with the production task employed. A meta-analysis indicates that in second language acquisition research, a wide range of production tasks is included in what is considered production-based instruction (i.e., anything from mechanical grammar drills to free production) without reference to vastly different task demands in these types of production<sup>7</sup>. This method, while balancing task demands, captures a key difference between speaking and listening in real life situations: the strict recall of information required for production, while comprehension only requires recognition.

This method could be used for comparing the effects of comprehension versus production training on different language phenomena in languages of interest. It has been successfully used to study grammatical agreement learning in an artificial language<sup>8</sup> and German gender agreement in noun phrases<sup>9</sup>. Because the materials of Hopman and MacDonald are publicly available at <https://osf.io/74kqe>, the protocol presented in this paper follows their implementation in PsychoPy,<sup>10</sup> but the basic contrast between the active production and comprehension trials presented here could be implemented for other languages, other grammatical phenomena, and using different experimental software.

In the experiment reported here, participants learn an artificial language that describes a visual world of various aliens moving around on several landscapes. A full sentence in this language consists of 7 words of different word types that always occur in a fixed word order (**Figure 1**). The full language consists of 18 content words and two function words (**Figure 2**), and the five content word categories each have 2–6 different words, leading to 432 different possible full sentences. Four words in each full sentence (i.e., determiner, color adjective, alien noun, verb) take the same suffix. There are two different types of aliens: scary-looking aliens, characterized by a multitude of eyes and legs, sharp teeth, and angular shapes; kind-looking aliens, characterized by a single big eye, two legs, a friendly smile, and rounded shapes. Note that for counterbalancing purposes,

the experiment program randomly assigns visual meanings to words within each category. One consistent mapping is used in the figures and demonstrations throughout this article, but another participant may learn for example, ‘saf’ means yellow and ‘fum’ means purple.

[Place **Figure 1** here]

[Place **Figure 2** here]

Participants can learn several grammatical regularities of interest in this language: First, there is the fixed word order. Second, the four (identical) suffixes that occur in a full sentence encode meaning in two different ways (see inset in **Figure 2**). The suffix encodes for plurality, denoting whether 1 or 2 aliens are present in the visual scene: short suffixes ‘ok’ and ‘us’ indicate singular; longer suffixes ‘oko’ and ‘usu’ indicate plural. The suffix also encodes for alien type, with ‘ok’ and ‘oko’ indicating that the alien in the image is scary-looking, and ‘us’ and ‘usu’ that it is kind-looking. Third, the exposure throughout the training paradigm is set up so that scary-looking aliens usually (on 83% of trials) occur with spotted patterns (freckles or patches) and only rarely (17% of trials) occur with striped patterns (curved or straight lines). The opposite is true for the kind-looking aliens; they usually occur with striped patterns and rarely with spotted patterns. This is also a language regularity because it creates a higher transition probability between, for example, the suffixes for scary-looking aliens and the words for spotted patterns. The experiment is set up so that there are no other co-occurrence regularities (e.g., each alien occurs equally often in each of the two colors throughout both training and testing).

During training, participants alternate between blocks of passive exposure trials and active training trials. During passive exposure trials, participants can learn about the language by simply watching images and listening to the language. During active training blocks, participants learn the language by actively practicing it in either active production or active comprehension training trials, depending on the condition they are in. The very first block of passive trials shows each of the 6 aliens once. Then, participants get an active block in which they practice with those same 6 black and white line drawings of the aliens. Gradually, new vocabulary is introduced, and the phrases get longer and the images more complicated (**Figure 3**). Throughout, each passive block is followed by an active block to practice with the same types of phrases and images, and production and comprehension participants learn the language in the exact same order with the exact same training structure; the correct target phrase and image pairing on active training trials is identical. Thus, the only difference between the two conditions is that production participants practice speaking the language on their active trials, whereas comprehension participants practice understanding the language on their active trials. After training, participants in both conditions get a set of identical comprehension tests that tap into their understanding of the vocabulary and the different grammatical regularities present in the alien language.

[Place **Figure 3** here]

**PROTOCOL:**

The following procedures were approved by the University of Wisconsin-Madison Social and Behavioral Science Institutional Review Board, and informed consent was obtained from each participant.

## 1. Materials

1.1. On the computer that will be used for the experiment, download the experiment files by going to <https://osf.io/74kqe/files/> in any browser.

1.1.1. At the top of the list, click on the parent folder called '**OSF Storage (United States)**'.

1.1.2. Click on the '**Download as zip**' button near the top of the screen to download the entire experiment.

1.2. Unpack all of the experiment materials by unzipping the downloaded .zip file.

1.2.1. Unpack '**slides.zip**'. The folder '**slides**' should now be populated with 11,520 image files ending in '.png'. If, when unzipping, Windows creates a subfolder of '**slides**' named '**slides**', copy all the image files from '**slides/slides**' into the main '**slides**' folder.

1.2.2. Open '**foldermaker.py**' in Psychopy (version 1.83.04) and click '**Run**'. The folder '**data**' should now be populated with subfolders '**s1**' through '**s399**', each of which should have subfolders called '**errortrials**' and '**recordings**'.

1.2.3. Open '**testgen.py**' in Psychopy and click '**Run**'. Open '**data/s1**' to check that it (and all other data subfolders) now has a file called '**fulltrialist.txt**'.

1.2.4. Open '**soundgen.py**' and click '**Run**'. The folder '**sounds/combined**' should now be populated with 2,057 sound files ending in '.wav'.

1.3. Prepare the experimental computer for participants.

1.3.1. Plug headphones to the computer.

1.3.2. Connect an external microphone to the computer.

1.3.3. Write '=' and '≠' on the sticky part of a post-it and use scissors to cut out the two symbols to the size of a keyboard key. Place the =-sticker on the 'L' key of the keyboard and the ≠-sticker on the 'F' key of the keyboard.

## 2. Artificial language learning experiment

2.1. Test whether the experiment works by running members of the research team through both conditions.

2.2. Open '**experiment.py**' in Psychopy and click '**Run**'. In the first pop-up, enter any subject number bigger than 1 to use for pilot participants. For example, in this test, 2 was used for the Comprehension Condition pilot participants and 3 for the Production Condition pilot participants.

NOTE: The number 1 is reserved for programming the experiment, and so will not run the full experiment. Only use this subject number if the code of the experiment will be changed.

2.3. In the second pop-up, enter the condition number. Entering '**1**' will run the Comprehension Training version of the experiment; entering '**2**' will run the Production Condition of the experiment.

NOTE: Steps 2.4–2.7 explaining the language training are also illustrated in the accompanying file '**Demo1training.pptx**'. Details (e.g., how long each image appears on the screen) are described in **Supplementary File 1**. While the protocol provided is detailed enough to understand the procedure and the results, it is best to review the supplemental materials for a more in-depth understanding of the method to run or adapt this protocol.

2.4. Perform a passive exposure trial by having participants in both conditions start learning the artificial language with the names of the 6 different aliens in 6 passive exposure trials (one per alien), where the participant is instructed to "listen to the language and watch the pictures on the screen".

2.5. Perform an active Comprehension trial (i.e., Comprehension Condition only). Prompt: Have the participant see an image on the screen and after 0.5 s the audio file with the target phrase is played. Response: Instruct the participant to "indicate by pressing the button whether the audio and the picture match or not" (i.e., '=' for match and '≠' for mismatch). The subject will then see a red cross if the response was incorrect and a green checkmark if it was correct. See **Figure 4** for an illustration of active comprehension trials with examples of correct and incorrect responses for both match and mismatch trials.

2.6. Perform the Active Production trial (i.e., Production Condition only). Prompt: Have the participant see the target image on the screen with a microphone icon below it. Response: Instruct the participant to "describe the picture out loud in the alien language" and press 'enter' to indicate that they are done speaking and save the microphone recording. See **Figure 4** for an illustration of active production trials with examples of a correct and incorrect response.

2.7. Perform correct pairing (identical for both conditions). In both conditions, have the participants receive the correct pairing right after responding on their own and instruct them to "pay attention to the correct pairing". Have participants see the target image and hear the target phrase that correctly describes it. Make it clear that a green square around the image indicates a correct pairing to ensure, participants in both conditions learn from the correct pairing irrespective of their own performance in the active task.

2.8. Refer to **Figure 3** and **Supplementary 1** for details about how participants learn the artificial language by alternating the passive and active training tasks described here to progress from single word to full sentence learning.

[Place **Figure 4** here.]

NOTE: Steps 2.9 and 2.10, which describe the different types of comprehension tests, are also illustrated in the accompanying file '**Demo2testing.pptx**'. Different types of trials (e.g., word order errors, grammatical agreement errors, correct sentences) are described in more detail in **Supplementary File 1**.

2.9. In a **Forced Choice Test** trial, have the participant see two different images of the same type on the left and right sides of the screen, with 'X' on the screen below the left picture and 'M' on the screen below the right picture. The participant will hear an auditory description that matches one of the two pictures. Instruct them to "indicate by pressing either 'M' or 'X' which of the two pictures you think matches the description". The buttons can be pressed before the end of the audio, and the trial ends immediately upon pressing the button.

NOTE: See **Figure 5A–C** for examples of three different types of forced choice test trials. There is a single word vocabulary test consisting of 18 items used as a prescreen, followed by the main forced choice test consisting of 66 trials total consisting of a mix of the three trial types illustrated in **Figure 5A–C**.

2.10. In **Error Monitoring Test** trials (i.e., a type of grammaticality judgement), have the participant hear a sentence and instruct them to "indicate whether the phrase is grammatical or not by pressing the appropriate button ('=' for grammatical, '≠' for ungrammatical) as fast and accurately as possible". Sentence presentation ends immediately on pressing the button. There are 124 error monitoring test trials total with three different types of trials, all of which are intermixed and played in a random order.

NOTE: See **Figure 5D–F** for illustrations of all different types of error monitoring test trials.

[Place **Figure 5** here.]

2.11. After running the two pilot participants, check that the data log saved and open a log in a spreadsheet program. For example, check that the data for participant 2 is in the folder '**s2**' and is called '**log2.txt**'. To open a .txt file in Excel, for example, right-click on **File Name|Open With|Excel**.

2.11.1. Scroll to the bottom and check that there is a separate line on the log file for each trial. Specifically, check that at the bottom of the file column 3 (**trialnr**) lists '**382**' for the final Error Monitoring trial.



2.11.2. Check that below the final trial, the log lists how many test trials of each type the participant got correct. For example, check that there is a number between 0 and 18 listed below 'total nr of correct vocabulary test trials was'.

### 3. Running the experiment

3.1. Recruit participants. In this study, 125 native English speakers were recruited from psychology extra credit research pool. Based on post-hoc power simulations (see Results), it is recommended to test at least 70 participants per condition, and more if an interaction between a learning condition and within-subject predictors (e.g., item type) are of interest.

3.2. Randomly assign participants to either the comprehension or the production condition, making sure to assign approximately the same number of participants to each condition.

3.3. Greet the participant and have them read and sign a consent form for the study.

3.4. Instruct participant to leave all items, including their phone, with the experimenter outside of the soundproof room with the experimental computer, headphones, and microphone.

3.5. Instruct the participant.

3.5.1. Tell them to put on the headphones and keep them on during the entire task.

3.5.2. Tell them that they will be learning a language and that all of the specific instructions will appear on the computer screen.

3.5.3. Point out which keys they will be using (i.e., the keys with '=' and '≠' stickers, 'enter' key, 'M' and 'X' keys) and that this information will also appear on the screen.

3.5.4. If the participant is in the Production condition, set out the microphone and tell them to speak into it when prompted.

3.6. Start the experiment by opening 'experiment.py' in Psychopy and clicking 'run'. Then, enter the appropriate subject number in the first pop-up and the condition number in the second pop-up. Then, tell the participant that they will be able to go through the experiment in a self-guided manner, as described in part 2 of this protocol.

3.7. When the participant is done with the experiment, thank them for participating and answer any questions they might have about the study.

### 4. Data processing

4.1. Trim the dataset according to prespecified criteria. Record how much data are removed and why.

4.1.1. If any participants behaved oddly or did not complete the experiment, remove their data. In this study, data from three participants who did not complete the experiment were removed.

4.1.2. If any participants did not meet a prespecified criterion, remove their data. For example, based on pilot testing for the present study set a criterion of at least 15 out of 18 correct responses on the single word vocabulary test was set.

4.1.3. Remove all trials where participants gave an incorrect response for reaction time analyses, because standard practice is to only analyze trials where participants give the correct response. Also remove outlier trials, defined in this study as trials in which a participant was slower than their own mean reaction time + three standard deviations. Finally, remove any trials with negative reaction time (if a more precise reaction time is calculated as in step 8.3 of **Supplementary File 1**). In this study, this left 78% of all test trial data to be used in the reaction time analysis.

#### **REPRESENTATIVE RESULTS:**

Average scores on the single word vocabulary test did not differ between the Production and Comprehension condition ( $t(120) < 1$ ). Data from 18 participants (8 comprehension, 10 production) who did not meet this criterion were removed. All further analyses reported here include data from the remaining 52 comprehension and 52 production participants. The results did not change when the data from these 18 participants were included.

Participants trained in the Production condition reliably outperformed participants in the Comprehension condition. Comprehension accuracy was analyzed using mixed effects logistic regression models<sup>11</sup>. Participants in the Production condition were significantly more accurate ( $P < 0.05$  using the likelihood ratio test<sup>12</sup>) than participants trained in the Comprehension condition on both the Suffix Understanding forced choice test trials (**Figure 6B**) and the Suffix Agreement Error monitoring test trials (**Figure 6E**). Production participants were also numerically more accurate than Comprehension participants on the other four tests (**Figure 6A,C,D,F**), though this difference was not statistically significant.

Reaction time data was also processed and analyzed following the steps outlined in **Supplementary File 1** under step 4.1.3 above to calculate a precise reaction time (**Figure 7**). Participants trained in the Production condition were significantly faster ( $P < 0.05$  using the Kenward-Roger approximation to the F-test<sup>13</sup>) than participants in the Comprehension condition on all three types of Forced Choice test trials (**Figure 7A,B,C**) and all three types of Error Monitoring test trials (**Figure 7D,E,F**), with small to medium effect sizes ( $d$  in range  $[-0.54, -0.19]$ )<sup>14,15,16</sup>. Based on post-hoc simulations, some of these tests were slightly underpowered (i.e., power in range [63%, 98%], average 73%).

All data, scripts with analyses, and regression model results reported here were reported previously in Hopman and MacDonald<sup>8</sup> and its supplementary materials. These analyses as well as the post-hoc power simulations and effect size calculations reported here for the first time can

be freely accessed online at <https://osf.io/bbf3c>.

#### FIGURE AND TABLE LEGENDS:

**Figure 1: Example sentence from the alien language.** This is an example of a full sentence in the alien language describing a video of the alien rising upwards. Below the sentence are the word types of all of the seven words in the sentence, a word-by-word translation into English and the full English translation as a sentence. The suffix 'ok' on four of the words in the sentence means this is an alien from the scary-looking (-sc) category, and that it is singular (-sg). Note that participants never see the language written out; they only hear it. This figure is adapted from Hopman and MacDonald<sup>8</sup>.

**Figure 2: Overview of the full alien language.** The full alien language has seven different word categories, with 1–6 words per category. The second row indicates, for each word category, how often each word from that category is practiced during passive (p) and active (a) training. For each of the 18 content words, the paired visual meaning is illustrated. For all 20 words, an English translation is given. Four of the 7 word categories (i.e., determiner, color adjective, alien noun, and verb) take a suffix, which is indicated by a '-' at the end of those words. In the bottom right corner, the suffixes, which express both plurality and alien type, are illustrated. This figure is adapted from the Supplementary Materials of Hopman and MacDonald<sup>8</sup>.

**Figure 3: Structure of the language training.** Every line in this figure represents a block of training trials, with the number of trials in the block as well as the type of training (i.e., active training or passive exposure) noted. For example, the very first block of six trials shows each of the 6 different aliens once. The green lines indicate blocks that introduce new vocabulary. The number of trials in these blocks was determined by the number of different stimuli in that vocabulary category. For example, the color vocabulary block had two trials because there were two color words. The blue lines indicate blocks in which participants practiced combining multiple aspects of the language. The number of trials in these blocks was always six because there was one trial per alien to make sure that participants got equal amounts of practice with all of the different aliens. The red lines indicate an extra active learning block that was added after pilot training to help participants learn the words whenever six new vocabulary words were introduced in one passive learning block (e.g., aliens singular and plural, verbs and landscapes). This figure is adapted from the Supplementary Materials of Hopman and MacDonald<sup>8</sup>.

**Figure 4: Active training trials.** All active training trials consisted of three parts: a prompt for the participant, the participant's response, and finally a correct pairing of the target image and the target phrase describing it. Note that the final part of the active learning trial, the correct pairing, was identical for all types of active learning trials. There were two types of active comprehension trials, in which the initial image matched the target phrase (match trials) and trials in which the initial image did not match the target phrase. For each trial type, examples of a correct response and an incorrect response by the participant are shown.

**Figure 5: Example test trials.** Example forced choice test and error monitoring test trials of different types. In forced choice test trials (A–C), the participants heard a phrase and saw two

images of the same type on the screen. By pressing a button they indicated which of the two images they believed was described by the auditory phrase. For each example forced choice trial, the critical word that allowed a participant to decide between the two pictures is printed in blue bold face. In error monitoring test trials (**D–F**) participants heard a full sentence and indicated by pressing a button whether or not they believed it was a grammatical sentence in the alien language. In the example trials with errors (**D,E**), the errors are printed in red bold face. In trials with grammatically correct sentences (e.g., **F**), participants did not know the correct answer until the final word of the sentence. For all different trial types, reaction time was counted at the start of the critical word that allowed the correct response, and this point in the phrase is indicated with an orange dotted line. Note that participants in this experiment never saw the language written out; they only heard these phrases. This figure is adapted from Hopman and MacDonald<sup>8</sup>.

**Figure 6: Representative accuracy results.** Model predictions for proportion correct on different comprehension tests of the artificial language. Numerically, Production participants were more accurate on all different tests than Comprehension participants, though the difference was only significant for two of the six different trial types. Error bars reflect 95% confidence intervals, asterisks indicate significant differences between learning conditions ( $P < 0.05$ ). Individual participants' vocabulary in phrases test scores were used as a covariate in all other tests. Post-hoc power estimates are reported for the learning condition predictor for each test based on simulations<sup>14</sup>. (**A**) Forced choice test, Vocabulary in Phrases trials. (**B**) Forced choice test, Suffix Understanding trials (post-hoc simulation = power 55%). (**C**) Forced choice test, Probabilistic Regularity trials. (**D**) Error monitoring test, Word Order Error trials. (**E**) Error monitoring test, Suffix Agreement error trials. (post-hoc simulation = power 100%). (**F**) Error monitoring test, Correct Sentence trials. Part of the data in this figure (**A,B,D,E**) was published earlier in Hopman and MacDonald<sup>8</sup>.

**Figure 7: Representative reaction time results.** Model predictions for reaction time on different comprehension tests of the artificial language. Production participants were significantly faster than Comprehension participants on all types of comprehension test trials. Error bars reflect 95% confidence intervals, asterisks indicate significant differences between learning conditions ( $P < 0.05$ ). The standardized effect size  $d$  for the learning condition predictor for each subfigure is reported in parentheses; note that negative effect sizes indicate that Production participants were faster than Comprehension participants. Post-hoc power estimates for the learning condition predictor for each test based on simulations are also reported<sup>15</sup>. Individual participants' vocabulary in phrases test scores were used as a covariate in all other tests. (**A**) Forced choice test, Vocabulary in Phrases trials ( $d = -0.19$ ; power 63%). (**B**) Forced choice test, Suffix Understanding trials ( $d = -0.54$ ; power 98%). (**C**) Forced choice test, Probabilistic Regularity trials ( $d = -0.22$ ; power 64%). (**D**) Error monitoring test, Word Order error trials ( $d = -0.28$ ; power 75%). (**E**) Error monitoring test, Suffix Agreement error trials ( $d = -0.33$ ; power 73%). (**F**) Error monitoring test, Correct Sentence trials ( $d = -0.25$ ; power 67%). Part of the data in this figure (**A,B,D,E**) was published earlier in Hopman and MacDonald<sup>8</sup>.

## DISCUSSION:

A procedure for studying the role of comprehension versus production practice in learning a

novel language is presented. As reported earlier in Hopman and MacDonald<sup>8</sup>, production-focused training results were found to be superior in learning an artificial language as compared to comprehension-focused training<sup>8</sup>. In follow-up research, there is accumulating evidence that production participants outperform comprehension participants in both comprehension and production accuracy<sup>9</sup>.

A critical component of the method presented here is the balancing of attention demands and amount of listening experience for active training tasks in both conditions (steps 2.5 and 2.6), as it allows for the analysis of possible underlying mechanisms that may lead to the benefits of production training. A likely possible mechanism for the production advantage seen is that planning an utterance for speaking requires the participant to maintain information in verbal working memory, thereby facilitating binding between the different elements of the utterance<sup>8</sup>. Because listening experience is also balanced between conditions, and there is still a significant benefit to the production condition, it is evident that the benefit of production is not just due to hearing oneself speak or increased task and attention demands typically associated with language production.

Earlier studies investigating language instruction found that comprehension practice, not production practice, leads to better performance when learning a new language<sup>17</sup>. While at first glance this seems to contradict the results of the present study, the methods are slightly different in that “production practice” in these older studies include passive repetition of a teacher’s input instead of a meaningful generation of one’s own utterance. Thus, another critical component of the protocol presented here is that participants are engaged in full language production; that is, they are required to recall words from long-term memory and create a sentence structure for the active production task. This is more focused on actual language production, which involves planning and structuring an utterance, than past studies, which have required participants to only repeat another person’s utterance (e.g., a teacher’s utterance).

Provided is a detailed guide (**Supplementary File 2**) on how to adapt this experiment for use with other stimuli (e.g., a different language). In principle, it is possible to apply the balanced active comprehension and production training (critical steps 2.5 and 2.6) as part of any language learning method. In practice, because the code implements a full training that depends on the number of words per category, the code that runs this full experiment (**‘experiment.py’**) only runs with languages with the exact same structure (e.g., an equal number of determiners, adjectives, nouns, etc.). However, this limitation is easily overcome. With the information presented in this manuscript, any experimental psychologist can implement this set of active production and comprehension training trials as part of any existing language learning training. For example, these same production and comprehension training trials were implemented in JavaScript in a differently structured artificial language learning experiment<sup>18,19</sup> and in Ibex Farm in a German learning experiment<sup>9,20</sup>.

Thus, future studies could use the balanced production and comprehension paradigm to investigate whether the benefits of production training found in this study hold true for other languages and other grammatical phenomena. Furthermore, the method could be adapted to

test other populations of learners, like children and people with learning disorders, to compare the usefulness of production and comprehension practice in different types of learners. The paradigm is flexible and the balanced contrast between active comprehension and production trials could be implemented as part of any existing language learning training.

#### **ACKNOWLEDGMENTS:**

We thank Teresa Turco for creating stimuli. We thank everyone in the Language and Cognitive Neuroscience Lab, Jenny Saffran, and Tim Rogers for helpful discussion. We thank Misty Kabasa for helpful feedback on an earlier draft of this manuscript. We thank Cassandra Jacobs for statistical consulting on effect size in mixed effects models. 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.

#### **AUTHOR CONTRIBUTIONS:**

EWMH and MCM created this method and conducted the original experiment. ML wrote the first draft of this paper under supervision of EWMH. EWMH rewrote the paper based on editor and reviewer suggestions. All authors provided feedback and edits on all submitted versions of the manuscript.

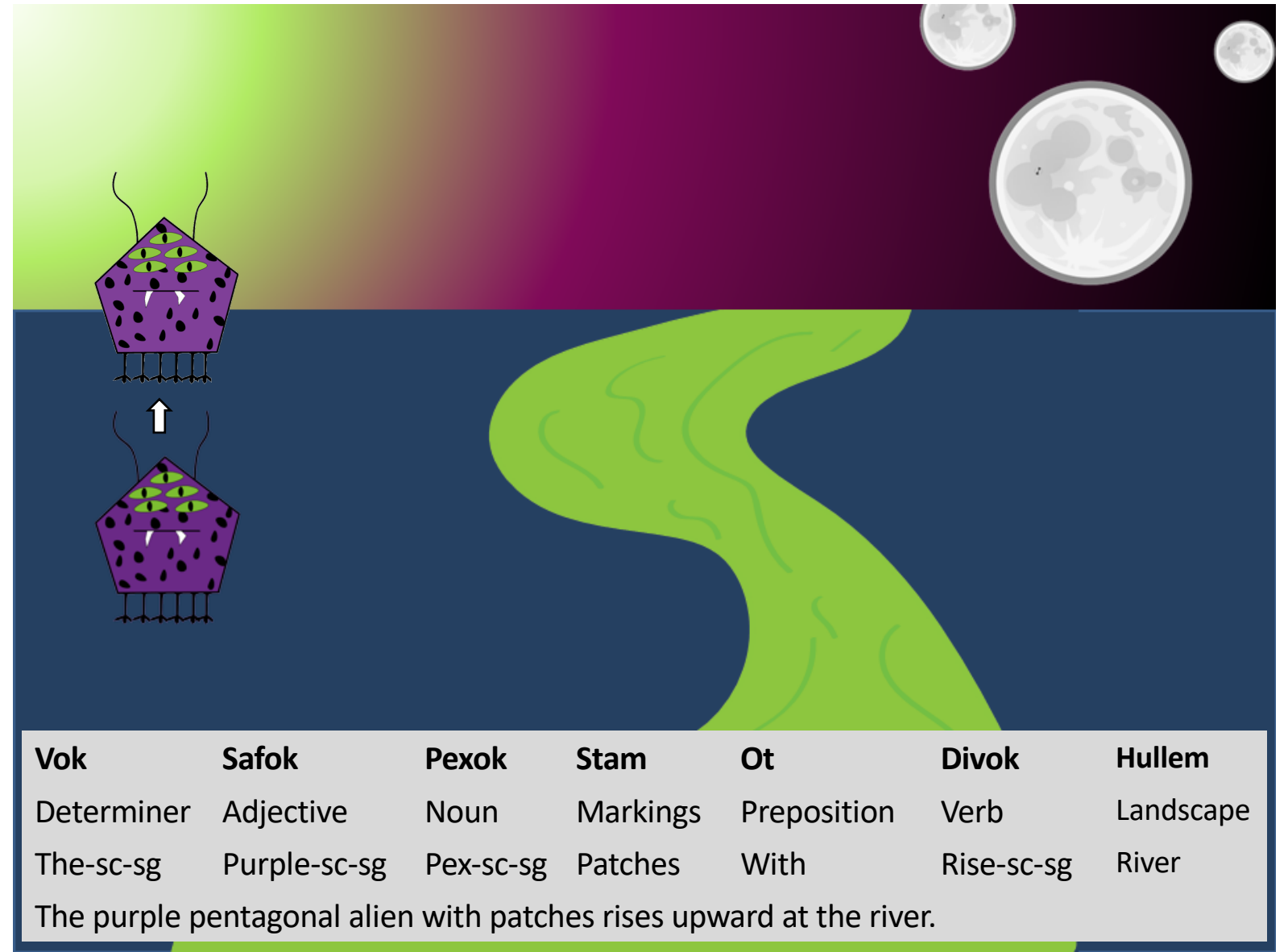
#### **DISCLOSURES:**

The authors have nothing to disclose.










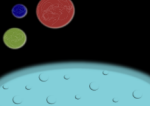
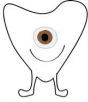







#### **REFERENCES:**

1. Krashen, S. D., Terrell, T. *The natural approach*. Alemany Press. Hayward, CA (1983).
2. Krashen, S. D. *Explorations in language acquisition and use*. Heinemann. Portsmouth, NH (2003).
3. MacLeod, C. M., Bodner, G. E. The production effect in memory. *Current Directions in Psychological Science*. **26** (4), 390-395 (2017).
4. Craik, F. I. M., Tulving, E. Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*. **104** (3), 268-294 (1975).
5. Carter, M. J., Ste-Marie, D. M. Not all choices are created equal: Task-relevant choices enhance motor learning compared to task-irrelevant choices. *Psychonomic Bulletin & Review*. **24** (6), 1879-1888 (2017).
6. Karpicke, J. D., Roediger, H. L., III. The critical importance of retrieval for learning. *Science*. **319** (5865), 966-968 (2008).
7. Shintani, N., Li, S., Ellis, R. Comprehension-based versus production-based grammar instruction: A meta-analysis of comparative studies. *Language Learning*. **63** (2), 296-329 (2013).
8. Hopman, E. W., MacDonald, M. C. Production practice during language learning improves comprehension. *Psychological Science*. **29** (6), 961-971 (2018).
9. Keppen, V., Hopman, E. W. M., Jackson, C. N. Production training benefits comprehension of grammatical gender in L2 German. *Talk presented at the International Symposium of Bilingualism*. Edmonton, AB (2019).
10. Peirce, J. W. PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*.

- 572 **162** (1-2), 8-13 (2007).
- 573 11. Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. Random effects structure for confirmatory  
574 hypothesis testing: Keep it maximal. *Journal of Memory and Language*. **68** (3), 255-279 (2013).
- 575 12. Molenberghs, G., Verbeke, G. Likelihood ratio, score, and Wald tests in a constrained  
576 parameter space. *The American Statistician*. **61** (1), 22-27 (2007).
- 577 13. Luke, S.G. Evaluating significance in linear mixed-effects models in R. *Behavior Research*  
578 *Methods*. **49** (4), 1494–1502 (2017).
- 579 14. Westfall, J., Kenny, D. A., Judd, C. M. Statistical power and optimal design in experiments in  
580 which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology:*  
581 *General*. **143** (5), 2020-2045 (2014).
- 582 15. Brysbaert, M., Stevens, M. Power analysis and effect size in mixed effects models: a tutorial.  
583 *Journal of Cognition*. **1** (1), 1-20 (2018).
- 584 16. Cohen, J. *Statistical power analysis for the behavioral sciences*, 2<sup>nd</sup> edition. Erlbaum. Hillsdale,  
585 NJ (1988).
- 586 17. Shintani, N., Ellis, R. The incidental acquisition of English plural -s by Japanese children in  
587 comprehension-based and production-based lessons: a process-product study. *Studies in Second*  
588 *Language Acquisition*. **32** (4), 607-637 (2010).
- 589 18. Hopman, E. W. M., Zettersten, M. Immediate feedback is critical for learning from your own  
590 productions. *Poster presented at Psycholinguistics in Flanders*. Ghent, Belgium (2018).
- 591 19. de Leeuw, J. R. jsPsych: A JavaScript library for creating behavioral experiments in a web  
592 browser. *Behavior Research Methods*. **47** (1), 1-12 (2015).
- 593 20. Drummond, A. Ibex Farm [software]. <http://spellout.net/ibexfarm/> (2013).





Determiner	Adjective	Noun	Markings	Pre-position	Verb	Landscape type
p: 72 a: 72	p: 28 a: 28	p: 12 a: 13	p: 13 a: 13	p: 39 a: 42	p: 13 a: 14	p: 13 a: 14
V-  the	Fum-  yellow	Teep-  tri-oval alien	Traw  curved lines	Ot  with	Div-  grow bigger	Kredel  mountainous
	Saf-  purple	Zout-  keyhole alien	Plim  straight lines		Pav-  move rightward	Chaftem  crater
		Weem-  heart alien	Chag  freckles		Zev-  rise upward	Hullem  river
		Mog-  kite alien	Stam  patches			
		Ket-  trapezoid alien				
		Pex-  pentagonal alien				




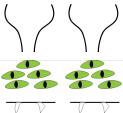

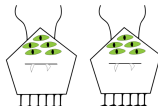





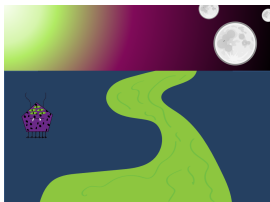
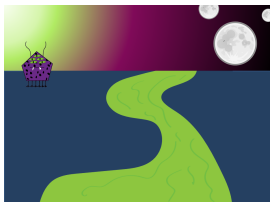

















































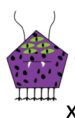
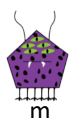



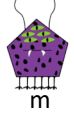

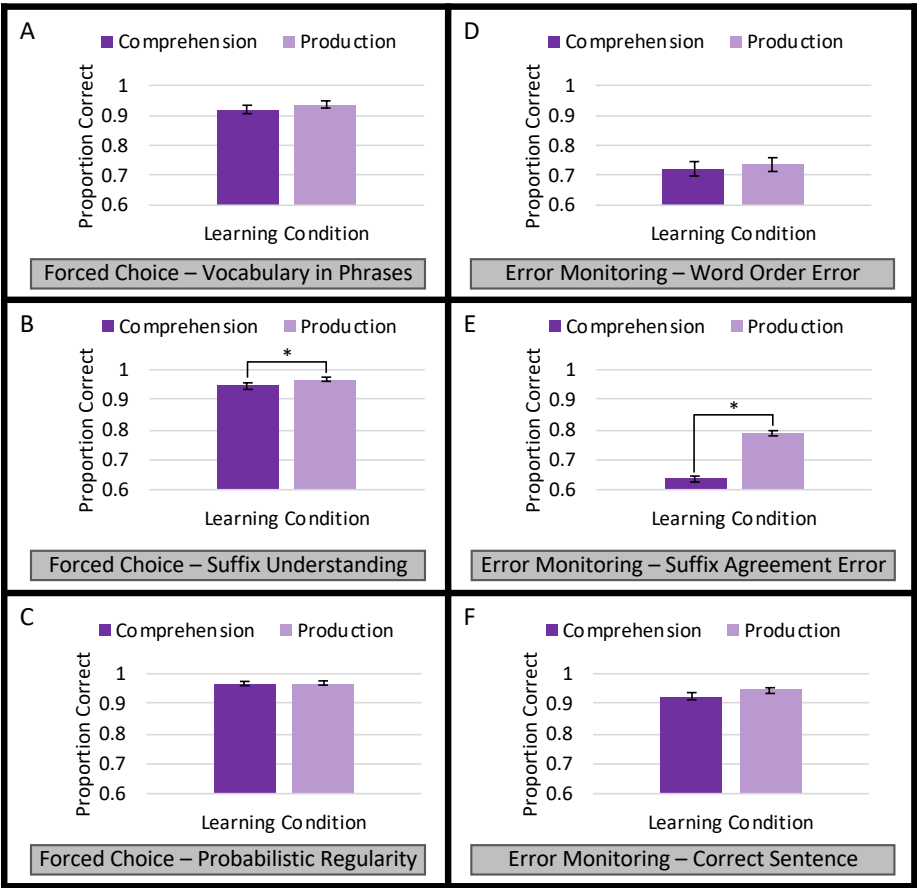
Suffixes	
-us  nice singular	-usu  nice plural
-ok  scary singular	-oko  scary plural

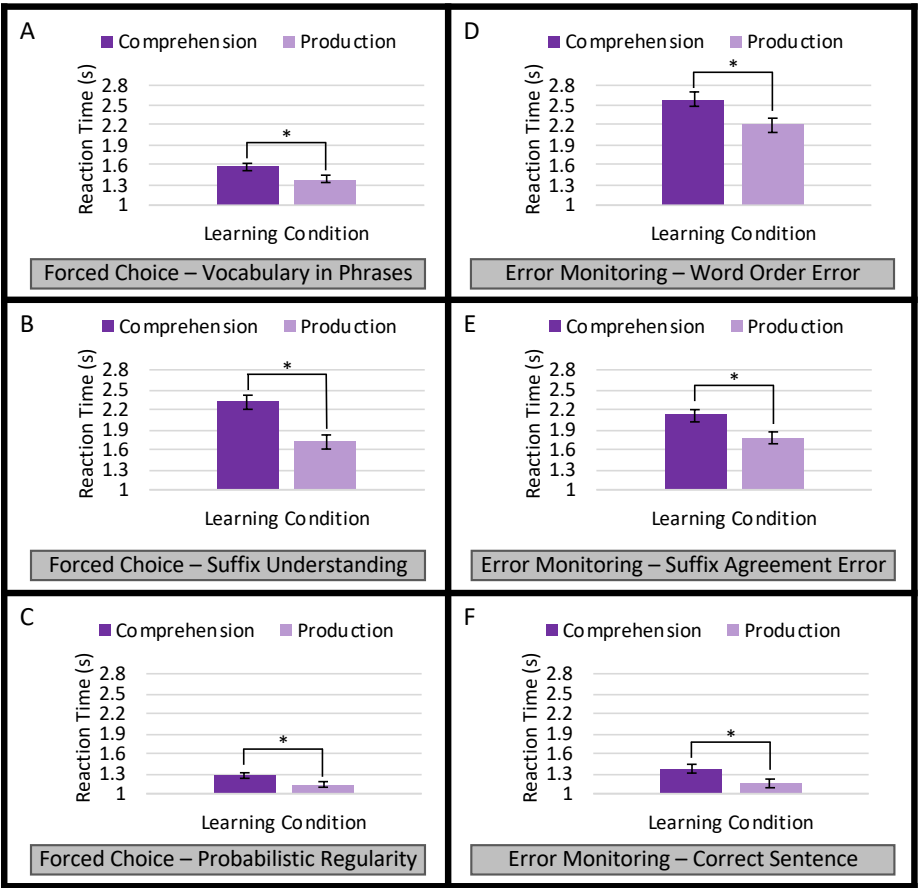
Figure 3

# of trials	Stimulus type	Block type	Example utterance, translation & image
6	singular monster vocabulary	passive exposure	"Vok Pexok" 
6	singular monster vocabulary	active learning	
6	singular monster vocabulary	active learning	
6	plural monster vocabulary	passive exposure	"Voko Pexoko" 
6	plural monster vocabulary	active learning	
6	plural monster vocabulary	active learning	
2	color vocabulary	passive exposure	"Saf" 
2	color vocabulary	active learning	
6	colored monster	passive exposure	"Vok Safok Pexok" 
6	colored monster	active learning	
4	markings vocabulary	passive exposure	"Stam" 
4	markings vocabulary	active learning	
6	colored monster with markings	passive exposure	"Vok Safok Pexok Stam Ot" 
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	"Hullem" 
3+ 3	verb and landscape vocabulary	active learning	
3+ 3	verb and landscape vocabulary	active learning	At the river.
6	full sentence	passive exposure	"Vok Safok Pexok Stam Ot Divok Hullem"   (only the first and last frame of the video are shown here)
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	
6	full sentence	passive exposure	
6	full sentence	active learning	
6	full sentence	active learning	
174	trials total in 31 blocks; 78 passive + <b>96 active</b> trials in 14 passive + <b>17 active</b> blocks		

			prompt	response	correct pairing
comprehension trial	match trial	correct response	 "vok pexok"  ≠  =	 ≠  = 	 "vok pexok" 
		incorrect response	 "vok pexok"  ≠  =	  ≠  =	 "vok pexok" 
	mismatch trial	correct response	 "vok pexok"  ≠  =	  ≠  =	 "vok pexok" 
		incorrect response	 "vok pexok"  ≠  =	  ≠  =	 "vok pexok" 
production trial		correct response	 	  "vok pexok"	 "vok pexok" 
		incorrect response	 	  "vus ketus"	 "vok pexok" 

<div>A</div> <div></div> <div>"vok fumok pexok stam ot"</div> <div></div> <div>x</div> <div></div> <div>m</div> <div>Forced Choice – Vocabulary in Phrases</div>	<div>D</div> <div></div> <div>"vok safok hullem pexok stam ot divok"</div> <div>≠</div> <div>=</div> <div>Error Monitoring – Word Order Error</div>
<div>B</div> <div></div> <div>"voko safoko pexoko  stam ot"</div> <div></div> <div>x</div> <div></div> <div>m</div> <div>Forced Choice – Suffix Understanding</div>	<div>E</div> <div></div> <div>"vok safok pexok stam ot divoko hullem"</div> <div>≠</div> <div>=</div> <div>Error Monitoring – Suffix Agreement Error</div>
<div>C</div> <div></div> <div>"vok safok pexok stam ot"</div> <div></div> <div>x</div> <div></div> <div>m</div> <div>Forced Choice – Probabilistic Regularity</div>	<div>F</div> <div></div> <div>"vok safok pexok stam ot divok hullem"</div> <div>≠</div> <div>=</div> <div>Error Monitoring – Correct Sentence</div>







Click here to access/download  
**Video or Animated Figure**  
Demo1Training.pptx





Click here to access/download  
**Video or Animated Figure**  
Demo2Testing.pptx



Name of Material/ Equipment	Company	Catalog Number
Browser	-	-
Desktop Computer	-	-
Experimental software	Psychopy	-
Headphones	LyxPro	-
Microphone	Blue	-
	Microsoft	
Software to open spreadsheets	Excel	-
Soundproof experiment room	-	-
Statistical analysis software	R	-
Stickers	Post-it	-

### **Comments/Description**

Use for downloading the experiment onto the computer.

Use for presenting the experiment on; use for analyzing data.

Psychopy version 1.83.04 is used for running the experiment, it is available on github.

Use for playing auditory stimuli to participants. Specifically, our lab currently uses HAS-10 over-ear open back studio headphones.

Use for recording production participants' training trials. Specifically, our lab uses Snowball microphones.

Use for a quick view of datalogs.

Use for running participants in.

Use for analyzing accuracy and reaction time data.

Use for marking keyboard keys used in the experiment.



Dear Dr. Steindel,

Thank you for inviting us to resubmit a second revision of our manuscript JoVE59946R1 titled **“A balanced test of comprehension versus production practice using artificial language learning”**. We would also like to thank the editor and the reviewers for providing us with a detailed list of suggested changes to improve our manuscript. Using this feedback, we have revised the manuscript extensively. Below, you will find the list of suggestions we received from the editor and the reviewers, with our responses interspersed in italics, and I have summarized the biggest changes here.

We have added two sets of supplemental materials. Per request of the editor, we moved detailed parts of the protocol without imperative steps out of the protocol and into supplemental materials titled ‘Supplement 1: Details of all tasks for implementing this method in a new script’. Per suggestion of both the editor and reviewer 2, we wrote ‘Supplement 2: Adapting the artificial language experiment’ to better enable other researchers to adapt our protocol. The OSF repository with our data and analyses (<https://osf.io/bbf3c/>) has been updated to reflect changes in this manuscript (e.g. two .Rmd scripts are added with power analyses and effect size calculations), and these updates are reflected in the file attached as a supplement to this manuscript titled ‘HopmanMacDonaldDataandAnalysisArchiveUpdated.zip’. One final change we made is in the author order. This change was agreed upon by all authors to reflect the work of the corresponding author on twice revising the manuscript. In order to clarify author contributions, we have added a small heading at the end of the manuscript that details the contributions each author made.

The revised manuscript still meets all journal guidelines on e.g. word length, number of paragraphs per section, etc.. We look forward to hearing your thoughts on this new version of our manuscript, and thank you again for this opportunity to resubmit the manuscript.

Yours sincerely, also on behalf of my co-authors,  
Elise Hopman

**From the editor’s email:**

Your manuscript, JoVE59946R1 "A balanced test of comprehension versus production practice using artificial language learning," has been editorially and peer reviewed, and the following comments need to be addressed. Note that editorial comments address both requirements for video production and formatting of the article for publication. Please track the changes within the manuscript to identify all of the edits.

*We have tracked changes in the manuscript so that all edits are clearly visible. In addition to the changes requested below, we have slightly shortened some existing parts of the introduction in order to stay within the 1500 word limit after adding a paragraph to it per reviewer 1's request. Please note that the line numbers are changed in this version due to changes to the manuscript.*

After revising and uploading your submission, please also upload a separate rebuttal document that addresses each of the editorial and peer review comments individually.

*This document responds to each comment individually.*

Please submit each figure as a vector image file to ensure high resolution throughout production: (.psd, ai, .eps., .svg). Please ensure that the image is 1920 x 1080 pixels or 300 dpi. Additionally, please upload tables as .xlsx files.

*We have uploaded all images as .pdf vector images as per the journal's current guidelines in 'Instructions\_for\_Authors.docx'. We have uploaded the only table (the table of materials) as a .xlsx file.*

#### **Editor's comments:**

1. Please take this opportunity to thoroughly proofread the manuscript to ensure that there are no spelling or grammar issues.

*We read through the manuscript and proofread it for spelling and grammar errors and edited it where needed.*

#### **Protocol:**

1. For each protocol step/substep, please ensure you answer the "how" question, i.e., how is the step performed? Alternatively, add references to published material specifying how to perform the protocol action. If revisions cause a step to have more than 2-3 actions and 4 sentences per step, please split into separate steps or substeps.

*- We have moved text from step 4.1.2. in the previous manuscript about pretest results to the representative results section because that step was 5 sentences and because it talked about results rather than a protocol step.*

*- We double checked and made sure every step in the protocol now has no more than 2-3 actions and 4 sentences per step and answers the 'how' question.*

#### **Specific Protocol steps:**

1. 2.4-2.12: Please rewrite section 2 so that it is in the imperative, per JoVE guidelines; it may be best to move some of it to a supplemental file as well.

*We took this advice and moved all steps without imperative instructions to a new supplement (Supplement 1: Details of all tasks for implementing this method in a new script). We kept essential steps for understanding the method in the main protocol (many of those were highlighted and already included imperatives). Now, all of the smaller details that are not*

*necessary to execute the protocol or understand the method (like precise timing of elements of each trial) are in the supplement. We believe this improves the clarity of the protocol greatly and improves focus on the most important aspects of the method.*

Figures:

1. Please obtain explicit copyright permission to reuse any figures from a previous publication. Explicit permission can be expressed in the form of a letter from the editor or a link to the editorial policy that allows re-prints. Please upload this information as a .doc or .docx file to your Editorial Manager account.

*We have requested and received explicit permission to reuse figures from our previous publication, and have uploaded the explicit permission in a file titled “HLMJoVESupplement3copyrightPermissionsFiguresSage.pdf” that is uploaded with our submission.*

2. Figure 6 is not currently cited in the text; it appears that the Figure 3 in the Results actually refers to Figure 6.

*Thank you for catching this error – it should indeed have said Figure 6 there instead of Figure 3. We have fixed this in the new manuscript.*

3. Figures 6 and 7: What statistical tests were used to obtain these p-values?

*We have added this information in the text now on lines 548 and 556-557. Furthermore, we have added references 12 and 13 that provide more information about these statistical tests.*

Discussion:

1. As we are a methods journal, please revise the Discussion to explicitly cover the following in detail in 3–6 paragraphs with citations:

a) Critical steps within the protocol

*The second paragraph in the discussion describes the critical aspects of our protocol, and we have added in text on line 662 to make explicit which steps of the protocol that text pertains to.*

b) Any modifications and troubleshooting of the technique

*We have included an extra supplement, titled: ‘Supplement 2: adapting the artificial language experiment’ that covers modifications in detail, and have addressed this, as well as the next comment in a new paragraph in the discussion section (lines 687-699). We have also uploaded a new document titled ‘TroubleshootingInstallation.docx’ to the OSF repository that this experiment can be downloaded from (<https://osf.io/74kqe/>) and that is already referenced in the manuscript on lines 124 and 188.*

c) Any limitations of the technique

*We have included a discussion of the particular situations in which this protocol in its entirety can be used versus situations in which our fully programmed experiment cannot be used in a new paragraph in the discussion section (lines 687-699). In those situations, the critical protocol steps (2.5. and 2.6.) could still be included as part of any existing language learning training, but would have to be implemented from scratch using the details in this manuscript. We have addressed this as a limitation.*

#### References:

1. Please ensure references have a consistent format.

*We have double checked the reference list and edited it as needed to meet the requirements in your 'instructions for authors'.*

#### Table of Materials:

1. Please ensure the Table of Materials has information on all materials and equipment used, especially those mentioned in the Protocol.

*We have added the microphone to the materials list, which was missing in our previous version, as well as the soundproof room and software for opening spreadsheets.*

#### **Reviewer 1:**

##### Minor Concerns:

- The introduction is clear and motivates the rationale for the development of this paradigm well. However, I would like to see a brief summary of some of the key design features of production and comprehension tasks typically used in this literature, and an explicit explanation of how the current task compares to those.

*Thank you for this great suggestion, we have included a new paragraph in the introduction that does what you ask (lines 104-113).*

- line 320: not clear what subtypes you're referring to here.

*The subtypes of trials that are meant are described in the following three sub-steps, and we have made that explicit in the text now. Note that all of this text has now moved to Supplement 1 per editor request, it is in section 5. of that supplement.*

- line 329: need more clarity on how the suffixes indicate both number and alien type, at a fairly early point in describing the task.

*We have included a more detailed explanation early in the manuscript (lines 147-150).*

- line 336: what are the participants actually asked to judge here?

*Thank you for pointing out that that was not clear in the previous write-up. We have added extra information here that we think clarifies this trial type. Note that all of this text has now moved to Supplement 1 per editor request, it is in section 5.3. of that supplement.*

- line 347: Why not call this 'grammaticality judgement'?

*We use the term 'error monitoring' for two reasons. The first is to be consistent with the results-oriented publication about this experiment (Hopman & MacDonald, 2018). Second, and this is the reason we chose to name it this way in the original article, typical grammaticality judgments with natural language stimuli are conducted on e.g. a 7-point scale where participants can indicate degrees of grammaticality. In natural languages, grammaticality is often not thought of as a black and white correct/incorrect judgment (e.g. Hammerly & Dillon, 2017 for an example with asymmetrical agreement attraction errors in English). Furthermore, in these tasks typically the participant is instructed to read the entire sentence before making a judgment. In our experiment on the other hand, participants were encouraged to make their judgment as soon as they knew whether the sentence contained an error or not, and auditory sentence presentation ended on button press. Because of these differences with typical grammaticality judgment tasks, we are sticking with the term 'error monitoring', but we have added some extra information to this step (both to indicate sentence presentation ends on button press and to indicate that this is, in fact, a type of grammaticality judgment; step 2.10., lines 407-413).*

- line 352: Is this a randomly assigned 'wrong spot' for the word?

*No, we had 4 different types of erroneous word orders. Since this is too detailed for the manuscript, we included this information with examples of each of the four different erroneous word orders in Table S1 in Supplement 1.*

- lines 357- 363: This section is hard to follow, without a clearer explanation of the suffixes earlier in the ms (e.g. around line 130).

*We have included a more detailed explanation around the place suggested (now lines 147-150).*

- line 439: How did you decide on 80% word-learning as your criterion?

*We set the threshold as being correct on at least 15 out of the 18 vocabulary items (83%), since for 8 pilot participants, their average score was 16 and only one scored below 15. We agree that 15 out of 18 is clearer than 80%, so we have rephrased that in the text (step 4.1.2., line 523).*

- line 557: Is this study reporting different results to Hopman and MacDonald? If so, clarify and make an explicit comparison with those.

*This study is reporting the same results as Hopman and MacDonald, with the addition of the probabilistic regularity test results that were presented as supplemental materials in Hopman and MacDonald (Figure 6C,F and Figure 7C,F in this manuscript). In the previous manuscript this was only made explicit in the Figure captions, we have added a phrase to the final sentence*

*of the representative results section (line 564) and to a sentence early in the discussion (line 656) to make this more explicit for readers.*

- Results: Please include the effect sizes for the group comparisons.

*As far as we know, the field is not completely in agreement yet about the correct way to calculate standardized effect sizes for single predictors in linear mixed effects models. After consulting with a statistics expert, we learned and applied the current best way to do so for linear mixed effects models. We have included these effect sizes for our Reaction Time analyses both in the results text (lines 559-560) and in the caption for figure 7 (with citations for the method we used to calculate them). We uploaded a new R script including these calculations to the OSF repository with the other analyses mentioned in this paper (AnalysisJoVEEffectSize.Rmd/.html). This file is also included in 'HopmanMacDonaldDataandAnalysisArchiveUpdated.zip', a supplement to this manuscript. Unfortunately, we have not been able to find any way to calculate effect sizes for the generalized linear mixed effects models we use to analyze accuracy – it is unclear whether the method used for linear mixed effects models generalizes to generalized linear mixed effects models.*

- Fig. 1: Please include a summary of the total number of blocks, as well as total number of passive vs active exposures to each element of the language.

*We have added the total number of trials and blocks, both overall and split up into passive and active, into Figure 3 (this is the same figure that was accidentally named Figure 1 in the previous submission). We have added the total number of active and passive exposures for each element of the language to Figure 2.*

## **Reviewer 2:**

I would also recommend some copy-editing with an eye to verb tense and subject-verb agreement-- I have listed a number of typos or rewording suggestions below.

*We have worked through the list of rewording suggestions below, and have re-read the entire manuscript and edited all verb tense and subject-verb agreement errors and inconsistencies we could find. Thank you for pointing this out.*

## **Minor Concerns:**

1. For readers who wish to extend the paradigm, it might be useful to add a short section describing how to change the properties of the artificial language for a new version of the script (which file(s) to alter and what dependencies they have). Similarly, it might be useful to describe how the visual stimuli were created (e.g., what was balanced in their creation, and what tools were used to generate them)-- or if this is in the original manuscript, to add a citation directing the reader to that.



*We agree that this would be helpful, and have added detailed instructions on this in a new supplement titled 'Supplement 2: Adapting the artificial language experiment'.*

2. How was power calculated for the experiment, if it was calculated? Add sample size recommendations to reader in section 3.1 if possible, based on power or previous precedent.

*We have included power based on post-hoc simulations (lines 560-561; captions figure 6 and 7) and a recommended number of participants as requested (step 3.1., lines 479-481). The script to run these analyses ('powerJoVE.Rmd/.html') is uploaded to our OSF repository with the other analyses and included in the supplement 'HopmanMacDonaldDataandAnalysisArchiveUpdated.zip'.*

3. Are there suggested headphone and microphone types (1.3.1 / 1.3.2)? If so, please add in.

*We have added the specific microphone and headphone types our lab uses to the table of materials.*

4. What is currently section 2.5.5 and 2.2.6 might fit under 2.5.4-- I suggest removing those section headings, and possibly doing some re-wording to eliminate redundancy.

*There is no section 2.2.6, but we assume 2.5.6. was meant, which does exist and is related to 2.5.4. and 2.5.5. We combined 2.5.5 and 2.5.6. into a single point (now 2.5. in Supplement 1).*

5. Figure 5: It seems like panel C has a different response than A and B? It might be worth adding a line at 521 describing the response for each of A, B, and C, to better make analogy to D, E, and F.

*The only trial type that has a different response is the one indicated in panel F; panel C has a response similar to panels A and B. We have added a sentence to the figure caption (lines 616-618, what used to be line 521) to make this clearer. We would be happy to clarify more if it is still unclear.*

Typos, rewordings, etc:

*Thank you for pointing out all of these, unless otherwise noted we have made the suggested changes.*

line 34: includes language comprehension and productive training tasks... (for clarity) 'productive' (effective) does not mean the same as 'production' (speaking), so we have kept 'production'.

line 67: ...which addressed in the method presented here

line 72: ...production and comprehension training<:> While...

line 78/79: ...the exact same language materials. (remove 'in a different way' as it is redundant) *We think it is important to emphasize this here, since this is at the core of the method.*

line 80: ...Each active training trial (remove s)

line 128: It is unclear that this is the same suffix on four words-- re-word to clarify.

line 199: Just use capital letters for both L and F.

*note: for consistency we have changed all indications of keys to capital letters.*

line 221: ...training version of the experiment<;> entering '2' will...

line 299: ...participants in both conditions take the same tests... (remove 'exact' as it is redundant)

line 318/319: ...the participant sees picture on side of the screen... (this wording made me think there were four pictures)

line 323: ...In this type of forced test trial... (remove s)

line 466: ...Production participants ...

line 545: ...Production participants ...

*Apologies, it is unclear what is meant here. In case it is the amount of white space around the words, that is automatically done in word because the text is justified to both edges of the page.*

## Supplement 1: Details of all tasks for implementing this method in a new script

Note that highlighted text is also present in the main manuscript, it is provided here as a frame of reference of the parts in the main manuscript that the details here correspond to. Numbered steps are numbered separately in this supplement and the main manuscript. For all highlighted steps, the corresponding step number in the protocol in the main manuscript is given in parentheses.

### Training tasks.

1. Participants in both conditions start learning the artificial language with the names of the 6 different aliens. This first happens in 6 passive exposure trials (1 per alien), where the participant is instructed to “listen to the language and watch the pictures on the screen”. (2.4 in manuscript)

1.1. Each passive exposure trial throughout the experiment has the same structure. The participant sees an image for 0.5 seconds. Then, with the picture still on the screen, audio starts to play to describe that image in the artificial language. The picture remains on the screen for another 1.5 seconds before a white screen shows for 0.5 seconds.

1.2. Then, the exposure is repeated: participants see the same picture for .5 seconds, listen to the audio file, and then see the picture for another 1 second.

1.3. Throughout passive exposure trials, the picture has a green box around it to indicate that participants can be sure it is a correct pairing of image and description.

2. After this first block of passive exposure, participants get active training with the same 6 alien stimuli. The active training trials are different for the two conditions (Figure 4).

2.1. Each individual active training trial consists of three parts: prompt, response and correct pairing. The prompt and response parts are different for production and comprehension participants. The correct pairing part is identical for production and comprehension participants.

**2.2. Active comprehension trial. Prompt:** The participant sees an image on the screen and after 0.5 seconds the audio file plays the target phrase. **Response:** The participant is instructed to “indicate with a button-press whether the audio and the picture match or not” (‘=’ for match and ‘≠’ for mismatch). The participant then sees a red cross if they were incorrect and a green checkmark if they were correct. (2.5 in manuscript)

**2.3. Active production trial. Prompt:** The participant sees the target image on the screen with a microphone icon below it. **Response:** The participant is instructed to “describe the picture out loud in the alien language”. The participant can press ‘enter’ to indicate that they are done speaking and save the microphone recording. (2.6 in manuscript)

**2.4. Correct pairing (identical for both conditions).** In both conditions, participants receive the

correct pairing right after making their own response. Participants see the target image and hear the target phrase that correctly describes it, with a green square around the image indicating that this is a correct pairing. This way, participants in both conditions can learn from the correct pairing, irrespective of their own performance in the active task. Participants are instructed to “pay attention to the correct pairing”. (2.7 in manuscript)

2.5. For comprehension participants, the target audio heard during the prompt is heard again during the correct pairing, accompanied by the target image. For match trials, this is the same picture that was shown during the prompt of the trial. For production participants, the target image seen during prompt and response is seen again during the correct pairing, accompanied by the target phrase.

2.6. In the first block of active training trials, participants in both conditions do 6 trials, 1 per alien.

3. After the first passive and active blocks of 6 trials each, participants are instructed on the screen that the rest of the training will consist of similar blocks of 2-6 passive trials, with each passive block followed by an active block to practice with the new words learned in the passive block. Throughout training, the images and the phrases describing them increase in difficulty.

3.1. The very first block of training shows simple, uncolored line drawings of aliens accompanied by a simple phrase consisting of a determiner and a noun.

3.2. As training builds up in difficulty, whenever a new visual property is added and a new word type is introduced into the phrase, a simple vocabulary block is first done (Figure 3), showing only that property. For example, a color vocabulary trial shows a colored square accompanied by a color word.

3.3. Then, a block is done that combines the new word type into a phrase with what was learned in earlier blocks. For example, the first combined block shows colored aliens that are described by a three-word phrase consisting of determiner, color adjective, alien noun.

3.4. The phrases and images build up in difficulty according to the block list shown in Figure 3. The final 12 blocks of training show a video of a colored, patterned alien moving against the backdrop of an alien landscape, accompanied by a full sentence audio consisting of seven words that describe the full scene. An example full sentence with its translation is shown in Figure 1.

**Tests after learning.** After training, participants in both conditions take the same tests to assess whether they’ve learned the language. Participants start with two forced choice tests and end with an error monitoring test.

4. The first forced choice test is a single word vocabulary test.

4.1. In a forced choice test trial, the participant sees two different images of the same type on

the left and right sides of the screen, with 'X' on the screen below the left picture and 'M' on the screen below the right picture (Figure 5). They hear an auditory description that matches one of the two pictures. They are instructed to "indicate by pressing either 'M' or 'X' which of the two pictures they think match the description". The buttons can be pressed before the end of the audio, and the trial ends immediately upon button-press. (2.9 in manuscript)

4.2. The single word vocabulary test consists of 18 trials, 1 for each content vocabulary word that has a visual depiction in Figure 2. For each trial, the foil picture is taken from the same category – for example, a trial testing the landscape word "hullem" would show both the image of the landscape with the river and one of the other landscape images. Whether the correct image (that matches the audio) is shown on the left or right side of the screen is randomly determined by the program.

5. The second forced choice test uses the same trial format: the participant sees two pictures, hears a phrase and indicates by button-press which of the two pictures matches the phrase. There are 66 trials total with 3 different subtypes of trials, all of which are intermixed and played in a random order. The 3 different subtypes are meant to test three different aspects of learning (vocabulary understanding in phrases, suffix understanding, and sensitivity to the probabilistic regularity), and are each described in turn in steps 2.9.1.-2.9.3..

5.1. Vocabulary in phrases (Figure 5A). In this type of forced choice test trials, vocabulary knowledge is tested. In order to do that, the foil image only differs from the target image in one vocabulary word – in the example in Figure 5a, the only difference between the images is the color. The word that is tested is the word that is different between the two images. There are 18 trials of this type, 1 testing each content word as part of a longer phrase.

5.2. Suffix understanding (Figure 5B). In this type of forced choice test trial, understanding of the suffixes on the noun phrase is tested. For example, to test understanding of singular versus plural, a phrase is heard and the type of alien described by that phrase is shown both on its own and in plural (Figure 5B). A participant can determine the correct answer if they understand that '-oko' and '-usu' are plural and '-ok' and '-us' are singular. There are 24 suffix understanding trials in total, 12 testing plurality and 12 testing alien type.

5.3. Probabilistic regularity (Figure 5C). In this subtype of trials, participants are tested for sensitivity to the probabilistic co-occurrence regularity that was present in the training trials – namely, that 83% of nice aliens had a striped pattern, and 83% of scary aliens had a spotted pattern. Participants hear a full noun phrase and see two aliens that only differ in their type of pattern (spotted versus striped; Figure 5C). There are 24 of these trials in total, 12 where the target image described by the phrase is a probable combination (nice/striped or scary/spotted, and 12 where the target image described by the phrase is an improbable combination (nice/spotted or scary/striped). If participants know that nice aliens tend to have striped patterns, they should be faster and more accurate to choose the correct image on probable trials, and slower and less accurate on improbable trials; however, participants can also choose the correct answer by simply waiting to hear the pattern word and choosing the image with

that pattern.

5.4. When the participant presses a button, the trial ends, even if the audio was not done playing yet. A white screen is shown for 0.5 seconds and then the next trial starts immediately. The program gives participants the chance to take a self-timed break after trials 22 and 44. At that point, participants can continue with the experiment by pressing 'enter'.

6. Error monitoring test. In error monitoring test trials (a type of grammaticality judgement), a participant hears a sentence and is instructed to "indicate by button press ('=' for grammatical, '≠' for ungrammatical) as fast and accurately as possible whether the phrase is grammatical or not". Sentence presentation ends immediately on button press. There are 124 trials total with 3 different types of trials, all of which are intermixed and played in a random order. (2.10 in manuscript)

6.1. Word order errors (Figure 5D). In this type of trial, one of the words in the sentence is moved to a different spot in the sentence. There are 32 of these trials in total. The critical word in these trials is the first word in the sentence that is in the wrong spot, like 'hullem' in the example (Figure 5D). Table S1 shows the four different word order errors. There are 8 trials with each of these 4 different word order errors.

Table S1. Example sentence illustrating the four different word order errors.

correct sentence	<i>Vusu Fumusu Teepusu Traw Ot Divusu Kredel</i>
word order error 1	<i>Vusu <b>Teepusu</b> Fumusu Traw Ot Divusu Kredel</i>
word order error 2	<i>Vusu Fumusu Teepusu Traw <b>Divusu</b> Ot Kredel</i>
word order error 3	<i>Vusu Teepusu Traw Ot Divusu <b>Fumusu</b> Kredel</i>
word order error 4	<i>Vusu Fumusu <b>Kredel</b> Teepusu Traw Ot Divusu</i>

*Note.* The boldfaced word in each erroneous sentence is the erroneously placed word.

6.2. Agreement errors (Figure 5E). In this type of trial, one of the 4 suffixes in the sentence is different from the other 3 suffixes in the sentence. There are  $2 \times 2 = 4$  different types of suffix errors: either the suffix on the alien word (adjacent to the rest of the noun phrase) or the suffix on the verb (non-adjacent to the noun phrase) is wrong; either the erroneous suffix indicates the wrong alien type or it indicates the wrong plurality. In the example trial, the suffix 'oko' on the verb differs from the suffix 'ok' on the determiner, adjective and noun (Figure 5E) – thus, this is an adjacent plurality error. There are 48 of these trials in total.

6.3. There are 44 grammatical sentences in this test (Figure 5F), to ensure that participants have to listen carefully. All sentences are novel, so even though all of the different words are familiar, during training participants have not encountered these sentences before.

6.4. When the participant presses a button, the trial ends. A white screen is shown for 0.5 seconds and then the next trial starts immediately. The program gives participants the chance to take a self-timed break after trials 25, 50, 75 and 100. During breaks, participants can continue with the experiment by pressing 'enter'.

7. When the experiment ends, participants see a brief description of the goal of the experiment on the screen and are told that they are welcome to ask the experimenter more questions.

8.1. For test trials, the column 'correctanswer?' indicates whether the participant got that test trial correct (1) or incorrect (0). This can be used for analyzing accuracy.

8.2. The column 'RT' records for each trial how long after the start of the audio the participant pressed a button.

8.3. This raw reaction time can be made more accurate by instead calculating reaction time starting at the onset of the critical word, which is defined as the first word that allows for a correct choice (orange dotted lines in Figure 5). Which word is the critical word in a sentence is indicated in the column 'critword'. Durations of all words in the trial are listed in the columns 'timew1' through 'timew7'. The accurate reaction time is calculated by subtracting the time between the start of the audio phrase and the start of the critical word from the raw reaction time for each trial.

NOTE: This may lead to negative reaction time on trials for which the participant made a button-press before they could possibly have known the answer. Remove these trials from the reaction time analysis, since they represent either guesses or misunderstandings.

Supplement 2: Adapting the artificial language experiment

**How to determine if the ‘experiment.py’ experimental code is usable for a different (artificial) language?**

It is possible to use our psychopy code, ‘experiment.py’, with different sound recordings and visual stimuli, but only if the new recordings and images have the exact same structure as the artificial language presented in our paper (Figure 2).

An example of an experiment where the ‘experiment.py’ code would be useful with newly recorded language stimuli would be a replication of this artificial language experiment with native speakers of another language than English – for example, Dutch. Since phonetics is explicitly not of interest in this experiment, in that case one would record a new version of the artificial language with the exact same structure, with each of the words and suffixes in Table S1 replaced by a pseudo-Dutch word: a non-word with Dutch pronunciation. This example will be used throughout this supplement. If desired, it would also be possible to create new visual referents, but again, the script expects visual referents that are structured into categories in the specific way we designed our visual world.

An example of an experiment where the ‘experiment.py’ code would not be useful would be a version of this experiment with a language that has an entirely different structure. For example, we ran a version of this experiment focused on agreement in German noun phrases<sup>9</sup>. Since German has three genders (masculine, feminine and neuter), whereas the artificial language only had two semantic types, we chose to program that experiment from scratch on a different platform that was commonly used in the lab where that experiment was run. Of course, we made sure to implement the balanced active comprehension and production trials that are central to the paradigm described in this paper. Furthermore, because we were teaching these participants a natural language, we created new images depicting real objects (e.g. a bike, a candle). The ‘experiment.py’ script randomizes assignment of words within a category to images (e.g. for each participant it is randomly determined whether the color ‘yellow’ is labeled ‘fum’ or ‘saf’ in the experiment), which is not compatible with a natural language, in which a word has the same meaning and visual referent for every participant.

Changing the structure of the language and visual world would result in needing to redo most of the script. In that case, only the functions in the script coding the trial types themselves could be used (the functions for passive exposure trials, active comprehension trials and active production trials). However, the main part of the script that codes the specific list of training blocks would not be useable. This is because the training (Figure 3, see also file ‘exposure.txt’) is programmed based on blocks that have 6 trials (because the language has 6 nouns), and based on having 2 colors that are balanced to appear equally often with each noun, and so on for each word type.

Table S1: the words in the artificial language used in Hopman & MacDonald (2018).

	t1	t2	t3	t4	t5	t6	t7
	Determiner	Color	Alien	Pattern	Preposition	Verb	Location
w1	V-	Fum-	Teep-	Traw	Ot	Div-	Kredel



w2		Saf-	Zout-	Plim		Pav-	Chaftem
w3			Weem-	Chag		Zev-	Hullem
w4			Mog-	Stam			
w5			Ket-				
w6			Pex-				

### Designing a new version the artificial language

- For each word type in Table S1, come up with an equally long list of pseudowords.
- Make sure that each word is pronounceable in the participants' native language. For example, the word 'traw' (t4w1) is not pronounceable in Dutch because it ends with 'aw', and so would have to be replaced.
- Make sure that each pseudoword is not a word (or almost the same as a word) in the participants' native language, or in a language participants are likely to be familiar with. For example, for Dutch participants the word 'zout' (t3w2) would not work, since this is the Dutch word for 'salt'. The word 'teep' pronounced using Dutch phonology sounds exactly like the English word 'tape' that most Dutch participants are familiar with, so it would also have to be replaced.
- Generate words within each word type to have roughly the same internal structure. For example, all location words in Table S1 have two syllables. All color words only have one syllable. And so on.
- Choose four suffixes so that, when each of the four suffixes is added on to each of the determiner, color, alien and verb words, the combination is still pronounceable and still not a real word. Depending on the research question, you may or may not want to make the suffixes similar in structure to the ones listed in Table S2.

Table S2: the suffixes in the artificial language used in Hopman & MacDonald (2018).

	n1	n2
s1	-us	-usu
s2	-ok	-oko

- Write out a list that has all possible words in the language. This list should thus have 5 entries for each of the color, alien and verb stems in Table S1. For example, the list would contain 'fum', 'fumok', 'fumoko', 'fumus' and 'fumusu'. Note that the determiner never occurs on its own without a suffix, whereas the color, alien and verb words do.
- Have this list read over by at least two native speakers from the same population as the participants, and have them point out any words that are either not pronounceable or sound too much like words they know.

- Replace any words that didn't meet the check in the previous step, and iterate this process until all words in the language are deemed both pronounceable and unlike real words.

- create a list with all of the words that need to be recorded in randomized order

- Enter the full list into one vertical column on an excel sheet
- In the next column, generate a list of random numbers
  - In the top cell, enter =rand()
    - Hit enter
    - A random number will appear
  - Drag the bottom corner of the cell down the column to generate a full list of random numbers
    - Make sure every word has a number next to it
  - Highlight the number column
  - Click on Sort & Filter
    - Select sort smallest to largest
    - Then select expand the selection
    - The random numbers will get ordered, and the words will be placed in a random order
- Copy and paste the random word list into a word document and print it off
- Repeat this procedure to generate a second random word list

### **Recording the Words:**

- Go to a soundproof recording room
- Plug in Snowball microphone
- Open Audacity
- Make sure that the recording device (specified next to the microphone picture on audacity tool bar) is 'Blue Snowball'
- Hit record to test the mic
- Once everything is set up, close all doors to the room
- Hit record
- Read the randomized word list
  - Enunciate
  - Speak slowly and clearly
  - Pause between each word
  - Make sure not to end words in a question (raising your voice)
  - If you mess up a word, pause then say the word again without pausing the recording
- Stop the recording (take care not to stop the recording too soon after you say the last word bc you will hear the click of the mouse in the recording)
- Save the audacity file
- Repeat the procedure with the second random word list

Splice up the Recordings and Save them as Sound Files:

- Next, you select each word and save it individually as a .wav sound file
- Open the audacity (.aup) file with a recorded word list
- Click and drag your mouse over a sound wave bubble
  - This should be one word
  - Make sure that you select the entire word and don't cut off any sounds
  - Do not leave too much room on either end of the word because this will create a silence in the recordings
- Hit Ctrl+C to copy the sound
- Hit Ctrl+N to open up a new audacity document
- Hit Ctrl+V to paste the word into the new audacity document
- Listen to the word the make sure you copied and pasted everything correctly
- Then select 'effects' from the options at the top of the screen
  - Hit 'normalize'
  - This will improve the sound quality
- Finally, go to save the file, go to file>export audio
  - Save the file as the word name followed by which list number (ex: teep1)
  - Then hit save
  - This will create a .wav sound file of the word
- Repeat this process for the entire audacity recording
  - You should end up with one file for every word
- Repeat this process for the other audacity recordings
  - Save the file as the word name followed by a number 2 (ex: teep2)

**Choose the Best Recording:**

- Create a separate folder to save all of the best sound files
- Listen to both recordings of a word (e.g. teep1.wav and teep2.wav)
- Choose the best version of the word (sounding most natural, without coughs or clicks, etc.)
- Copy this sound file into the separate folder
- Repeat for all words in the language
  - It's easiest to search in your saved files for the word root (ex: teep) then all of the endings for that word (teepok, teepoko, teepus, teepusu) will also appear and you won't accidentally forget a word
  - Double check that you have a sound file for each word in the language

**Save the Recordings in the format Psychopy expects them:**

- In order for the experiment to recognize the word files and generate sentences, the sound files need to be saved in a systematic code
- Our code specifies word type (t), word number within a word type (w), semantic type (s), and number (n)

- Word type specifies the different kinds of words (i.e.: determiner, color, alien, pattern, preposition, verb, or location)
- Word number within the type differentiates between the words within a specific type (i.e.: color #1 v. color #2)
- Semantic type differentiates between word endings (i.e. scary v. nice/ us v. ok)
- Number specifies whether the word is plural or not (i.e. ok v. oko)
- The code will have each letter (t,w,s, and n) followed by a number
  - Ex: Fumoko (color word 1 with semantic type 2 and plural)
    - t2w1s2n2
- go to the folder with all of the best sound files
- rename all of the files in code
- Because the computer will reorder the files alphabetically as you rename them, it may be easiest to upload the files to box, then rename them. Box keeps the files in the order they were uploaded.
  - That way you can look at the word above to copy the format (ie teepok has a similar code to teepoko)

**\*\***There are a lot of little ways to make errors during this process. Try to double check your work once you've finished to make sure that you have a sound file for each word and that all of the words are named correctly. It's advisable to have another person double check this as well.

Using the sound files recorded in this way, run the 'soundgen.py' file as explained in the main manuscript (step 1.2.4. in the protocol) to generate longer phrases using the newly recorded version of the artificial language. The experiment should otherwise run exactly as specified in the protocol.

### **Creating the visual stimuli**

- once the language is designed, create visual stimuli to depict every possible phrase and sentence in the language.

- in our case, we started with creating the images for vocabulary words, that are shown in miniature form in Figure 3.

- We first created line drawings of 8 aliens from scratch in adobe illustrator. Since we have two semantic types, we created two sets of aliens with distinct visual properties. One set has a single eye, two legs, rounded shapes and a smile. The other set has five small eyes, multiple wiry legs, antenna and two teeth showing. Note that we ended up using only 6 of these 8 aliens in the experiment described here.

- We then created 4 different possible skin patterns for the aliens, in groups of 2 (there are 2 different spotted patterns and 2 different striped patterns). This was done so that we could implement a probabilistic regularity, where each type of alien was associated mostly with one type of pattern (e.g. most of the scary aliens seen during training had spotted patterns). We started with creating 4 squares showing just the black patterns on a white background (Figure 3, pattern column).

- We then created versions of each of the 8 different aliens in Adobe Illustrator with each of the 4 possible skin patterns, giving us  $4 \times 8 = 32$  different patterned alien pictures.
- We then chose 6 different colors the aliens could have. We created 6 different squares showing just that color.
- We also created versions of all of the aliens, both patterned and unpatterned, in each color. This led to  $6 \text{ (colors)} \times 5 \text{ (4 patterns + unpatterned)} \times 8 = 240$  different pictures of colored, (un)patterned aliens.
- At this point in stimuli creation, we piloted training up until full noun phrases (describing colored, patterned aliens) and realized that 6 different alien nouns and 6 different color words would take too long for participants to master, so we chose 2 of the 6 colors to work with for this experiment.
- We created 3 different alien landscapes in powerpoint, making sure each landscape was visually distinct, not too cluttered to take away attention from the aliens and roughly equally interesting to look at. We also made sure each landscape would clearly show aliens in our 2 colors without the alien fading into the background.
- We created 3 different verbs that we were able to visually depict as aliens moving without the actual alien picture needing to change. We chose the verbs to have different end states, so that at the end of the movement it would still be possible to see what movement had taken place. We chose growing (the alien becomes bigger in 10 steps), moving to the right (the alien moves to the right in 10 steps) and rising (the alien rises up in 10 steps).
- for each verb, we first created 10 frames that show a black rectangle doing that action against a white background (moving to the right, growing, rising up).
- we then created frames showing every possible verb with every possible patterned alien performing each action against each landscape, both on its own and with two of the same aliens adjacent to each other for depicting the plural. We did this by pasting the patterned alien(s) into the landscape either in slightly different positions for each frame (for moving to the right and rising up) or pasting in a slightly enlarged picture of that alien for each frame (for growing). Thus, we created  $10 \text{ (frames)} \times 2 \text{ (singular/plural)} \times 3 \text{ (landscape)} \times 3 \text{ (verb)} \times 4 \text{ (pattern)} \times 2 \text{ (color)} \times 8 \text{ (alien)} = 11520$  different frames. All previous steps were done by hand using a combination of powerpoint and illustrator, this step was done using a matlab script for both efficiency and precision.
- for our script to be able to read in the stimuli, use the naming conventions detailed in file 'FileNames for stimuli explained explained.docx' that can be found on OSF at <https://osf.io/74kqe/>

Friday, August 23, 2019 at 11:46:56 AM Central Daylight Time

---

**Subject:** RE: Request for reuse of picture used in a psychological science article I authored  
**Date:** Friday, August 9, 2019 at 12:35:38 PM Central Daylight Time  
**From:** permissions (US)  
**To:** Elise Hopman

Dear Elise,

You may adapt the figures but please indicated in the credit that the figure is adapted.

Kind regards,

Mary Ann Price  
*Rights Coordinator*  
SAGE Publishing  
2600 Virginia Ave NW, Suite 600  
Washington, DC 20037  
USA

T: 202-729-1403  
[www.sagepublishing.com](http://www.sagepublishing.com)

---

**From:** Elise Hopman <[hopman@wisc.edu](mailto:hopman@wisc.edu)>  
**Sent:** Tuesday, July 30, 2019 5:03 PM  
**To:** permissions (US) <[permissions@sagepub.com](mailto:permissions@sagepub.com)>  
**Subject:** Re: Request for reuse of picture used in a psychological science article I authored

Dear Mary Ann Price,

Thank you for the permission to reuse the three figures from the article.

Could you also clarify whether I have permission to reuse (adjusted) versions of the figures in the supplementary online materials?

I am asking this because I ended up redrawing all figures again, the only ones in the new manuscript that look similar to the ones in the Psych Science paper are versions of Figure 5 from the main manuscript and Figures S1 and S2.

For a different outreach project I am reusing an adjusted version of Figure 2 from the Psychological Science paper.

Thank you,  
Elise Hopman

---

**From:** "permissions (US)" <[permissions@sagepub.com](mailto:permissions@sagepub.com)>  
**Date:** Friday, July 26, 2019 at 3:25 PM  
**To:** Elise Hopman <[hopman@wisc.edu](mailto:hopman@wisc.edu)>

**Subject:** RE: Request for reuse of picture used in a psychological science article I authored

Dear Elise Hopman,

Thank you for your request. In regards to the 3 figures from the article I am pleased to report that in this instance we can grant your request without a fee.

**Please accept this email as permission for your request as you've detailed below. Permission is granted for the life of the edition on a non-exclusive basis, in the English language, throughout the world in all formats provided full citation is made to the original SAGE publication. Permission does not include any third-party material found within the work.**

In regards to the supplemental material, you retain the copyright and exclusive rights to that material so you do not need our permission.

If you have any questions, or if we may be of further assistance, please let us know.

Kind regards,

Mary Ann Price  
Rights Coordinator  
SAGE Publishing  
2600 Virginia Ave NW, Suite 600  
Washington, DC 20037  
USA

T: 202-729-1403  
[www.sagepublishing.com](http://www.sagepublishing.com)

Los Angeles | London | New Delhi  
Singapore | Washington DC | Melbourne

---

**From:** Elise Hopman <[hopman@wisc.edu](mailto:hopman@wisc.edu)>  
**Sent:** Monday, July 15, 2019 1:48 PM  
**To:** permissions (US) <[permissions@sagepub.com](mailto:permissions@sagepub.com)>  
**Subject:** Request for reuse of picture used in a psychological science article I authored

Dear Sage Publishing,

I am writing to ask permission to reuse some of the figures in an article I published with you in two new articles I am hoping to publish. Here is a link to the original article I published in Psychological Science, a sage journal: <https://journals.sagepub.com/doi/abs/10.1177/0956797618754486?journalCode=pssa#articlePermissionsContainer>

I am hoping to use figures 2, 4 and 5, out of the main text, as well as figure S1 and table S1 of the supplemental materials for a methods manuscript I am planning to submit to another (non-SAGE) journal. Is this allowed? I read your guidelines for sage authors here (<https://us.sagepub.com/en-us/nam/journal-author-archiving-policies-and-re-use>), but that page only seems to deal with the article as a whole, not with individual pictures. Of course, I would cite the original source in the figure captions in the new publication.

If I am not allowed to reuse these exact figures for free, then I was wondering whether I am allowed to use different versions that convey the same information. For example, I could render new graphs of part of the information in Figure 5 in the original manuscript (see the two attachments for the figure submitted to Psychological Science and my new rendering of it) – is that allowed?

Thank you,  
Elise Hopman  
<http://lcnl.wisc.edu/elise-hopman/>



## Elise W.M. Hopman

Elise Hopman is currently a graduate student in the department of Psychology at the University of Wisconsin – Madison, USA. She is interested in studying language learning, and her research in Maryellen MacDonald’s lab focuses on the relationship between language production and language comprehension during language learning.

Elise is originally from the Netherlands, where she obtained bachelor’s degrees in both Mathematics and Physics, as well as a master’s degree in Cognitive Neuroscience (Language specialization), all at the Radboud University in Nijmegen.


## Mackenzie Ludin

Mackenzie Ludin graduated from the University of Wisconsin – Madison, USA in December 2018 with a BS in Psychology, and is currently pursuing a master's degree in Occupational Therapy. During her Bachelor’s degree, she was a student Research Assistant in the Language and Cognitive Neuroscience Lab in the Psychology Department.

## Maryellen C. MacDonald

Maryellen MacDonald is Donald P. Hayes professor of Psychology at the University of Wisconsin-Madison. Her research investigates the links between language comprehension, language production, and memory representations. Although these fields are often studied independently, her work focuses on the intrinsic connections among them.

Maryellen MacDonald received her B.A. from the University of Texas in an interdisciplinary program combining cognitive psychology, linguistics, and some computer science. She received her Ph.D. in Psychology (Linguistics minor) from UCLA in 1986. Before landing at the University of Wisconsin-Madison, Maryellen did a postdoc at Carnegie Mellon University and held faculty positions at Northeastern, MIT, and the University of Southern California.



Click here to access/download  
**Supplemental Coding Files**

**HMDataandAnalysisArchiveUpdated.zip**



## ARTICLE AND VIDEO LICENSE AGREEMENT

Title of Article:

Training language comprehension versus production in a balanced way

Author(s):

Mackenzie Ludin, Elise W.M. Hopman, Maryellen C. MacDonald

Item 1: The Author elects to have the Materials be made available (as described at <http://www.jove.com/publish>) via:

☐

Standard Access

☒

Open Access

Item 2: Please select one of the following items:

☒

The Author is **NOT** a United States government employee.

☐

The Author is a United States government employee and the Materials were prepared in the course of his or her duties as a United States government employee.

☐

The Author is a United States government employee but the Materials were NOT prepared in the course of his or her duties as a United States government employee.

### ARTICLE AND VIDEO LICENSE AGREEMENT

1. **Defined Terms.** As used in this Article and Video License Agreement, the following terms shall have the following meanings: “**Agreement**” means this Article and Video License Agreement; “**Article**” means the article specified on the last page of this Agreement, including any associated materials such as texts, figures, tables, artwork, abstracts, or summaries contained therein; “**Author**” means the author who is a signatory to this Agreement; “**Collective Work**” means a work, such as a periodical issue, anthology or encyclopedia, in which the Materials in their entirety in unmodified form, along with a number of other contributions, constituting separate and independent works in themselves, are assembled into a collective whole; “**CRC License**” means the Creative Commons Attribution-Non Commercial-No Derivs 3.0 Unported Agreement, the terms and conditions of which can be found at: <http://creativecommons.org/licenses/by-nc-nd/3.0/legalcode>; “**Derivative Work**” means a work based upon the Materials or upon the Materials and other pre-existing works, such as a translation, musical arrangement, dramatization, fictionalization, motion picture version, sound recording, art reproduction, abridgment, condensation, or any other form in which the Materials may be recast, transformed, or adapted; “**Institution**” means the institution, listed on the last page of this Agreement, by which the Author was employed at the time of the creation of the Materials; “**JoVE**” means MyJoVE Corporation, a Massachusetts corporation and the publisher of The Journal of Visualized Experiments; “**Materials**” means the Article and / or the Video; “**Parties**” means the Author and JoVE; “**Video**” means any video(s) made by the Author, alone or in conjunction with any other parties, or by JoVE or its affiliates or agents, individually or in collaboration with the Author or any other parties, incorporating all or any portion

of the Article, and in which the Author may or may not appear.

2. **Background.** The Author, who is the author of the Article, in order to ensure the dissemination and protection of the Article, desires to have the JoVE publish the Article and create and transmit videos based on the Article. In furtherance of such goals, the Parties desire to memorialize in this Agreement the respective rights of each Party in and to the Article and the Video.

3. **Grant of Rights in Article.** In consideration of JoVE agreeing to publish the Article, the Author hereby grants to JoVE, subject to **Sections 4** and **7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Article in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Article into other languages, create adaptations, summaries or extracts of the Article or other Derivative Works (including, without limitation, the Video) or Collective Works based on all or any portion of the Article and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. If the “Open Access” box has been checked in **Item 1** above, JoVE and the Author hereby grant to the public all such rights in the Article as provided in, but subject to all limitations and requirements set forth in, the CRC License.

## ARTICLE AND VIDEO LICENSE AGREEMENT

4. **Retention of Rights in Article.** Notwithstanding the exclusive license granted to JoVE in **Section 3** above, the Author shall, with respect to the Article, retain the non-exclusive right to use all or part of the Article for the non-commercial purpose of giving lectures, presentations or teaching classes, and to post a copy of the Article on the Institution's website or the Author's personal website, in each case provided that a link to the Article on the JoVE website is provided and notice of JoVE's copyright in the Article is included. All non-copyright intellectual property rights in and to the Article, such as patent rights, shall remain with the Author.

5. **Grant of Rights in Video – Standard Access.** This **Section 5** applies if the "Standard Access" box has been checked in **Item 1** above or if no box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby acknowledges and agrees that, Subject to **Section 7** below, JoVE is and shall be the sole and exclusive owner of all rights of any nature, including, without limitation, all copyrights, in and to the Video. To the extent that, by law, the Author is deemed, now or at any time in the future, to have any rights of any nature in or to the Video, the Author hereby disclaims all such rights and transfers all such rights to JoVE.

6. **Grant of Rights in Video – Open Access.** This **Section 6** applies only if the "Open Access" box has been checked in **Item 1** above. In consideration of JoVE agreeing to produce, display or otherwise assist with the Video, the Author hereby grants to JoVE, subject to **Section 7** below, the exclusive, royalty-free, perpetual (for the full term of copyright in the Article, including any extensions thereto) license (a) to publish, reproduce, distribute, display and store the Video in all forms, formats and media whether now known or hereafter developed (including without limitation in print, digital and electronic form) throughout the world, (b) to translate the Video into other languages, create adaptations, summaries or extracts of the Video or other Derivative Works or Collective Works based on all or any portion of the Video and exercise all of the rights set forth in (a) above in such translations, adaptations, summaries, extracts, Derivative Works or Collective Works and (c) to license others to do any or all of the above. The foregoing rights may be exercised in all media and formats, whether now known or hereafter devised, and include the right to make such modifications as are technically necessary to exercise the rights in other media and formats. For any Video to which this **Section 6** is applicable, JoVE and the Author hereby grant to the public all such rights in the Video as provided in, but subject to all limitations and requirements set forth in, the CRC License.

7. **Government Employees.** If the Author is a United States government employee and the Article was prepared in the course of his or her duties as a United States government employee, as indicated in **Item 2** above, and any of the licenses or grants granted by the Author hereunder exceed the scope of the 17 U.S.C. 403, then the rights granted hereunder shall be limited to the maximum

rights permitted under such statute. In such case, all provisions contained herein that are not in conflict with such statute shall remain in full force and effect, and all provisions contained herein that do so conflict shall be deemed to be amended so as to provide to JoVE the maximum rights permissible within such statute.

8. **Protection of the Work.** The Author(s) authorize JoVE to take steps in the Author(s) name and on their behalf if JoVE believes some third party could be infringing or might infringe the copyright of either the Author's Article and/or Video.

9. **Likeness, Privacy, Personality.** The Author hereby grants JoVE the right to use the Author's name, voice, likeness, picture, photograph, image, biography and performance in any way, commercial or otherwise, in connection with the Materials and the sale, promotion and distribution thereof. The Author hereby waives any and all rights he or she may have, relating to his or her appearance in the Video or otherwise relating to the Materials, under all applicable privacy, likeness, personality or similar laws.

10. **Author Warranties.** The Author represents and warrants that the Article is original, that it has not been published, that the copyright interest is owned by the Author (or, if more than one author is listed at the beginning of this Agreement, by such authors collectively) and has not been assigned, licensed, or otherwise transferred to any other party. The Author represents and warrants that the author(s) listed at the top of this Agreement are the only authors of the Materials. If more than one author is listed at the top of this Agreement and if any such author has not entered into a separate Article and Video License Agreement with JoVE relating to the Materials, the Author represents and warrants that the Author has been authorized by each of the other such authors to execute this Agreement on his or her behalf and to bind him or her with respect to the terms of this Agreement as if each of them had been a party hereto as an Author. The Author warrants that the use, reproduction, distribution, public or private performance or display, and/or modification of all or any portion of the Materials does not and will not violate, infringe and/or misappropriate the patent, trademark, intellectual property or other rights of any third party. The Author represents and warrants that it has and will continue to comply with all government, institutional and other regulations, including, without limitation all institutional, laboratory, hospital, ethical, human and animal treatment, privacy, and all other rules, regulations, laws, procedures or guidelines, applicable to the Materials, and that all research involving human and animal subjects has been approved by the Author's relevant institutional review board.

11. **JoVE Discretion.** If the Author requests the assistance of JoVE in producing the Video in the Author's facility, the Author shall ensure that the presence of JoVE employees, agents or independent contractors is in accordance with the relevant regulations of the Author's institution. If more than one author is listed at the beginning of this Agreement, JoVE may, in its sole

## ARTICLE AND VIDEO LICENSE AGREEMENT

discretion, elect not take any action with respect to the Article until such time as it has received complete, executed Article and Video License Agreements from each such author. JoVE reserves the right, in its absolute and sole discretion and without giving any reason therefore, to accept or decline any work submitted to JoVE. JoVE and its employees, agents and independent contractors shall have full, unfettered access to the facilities of the Author or of the Author's institution as necessary to make the Video, whether actually published or not. JoVE has sole discretion as to the method of making and publishing the Materials, including, without limitation, to all decisions regarding editing, lighting, filming, timing of publication, if any, length, quality, content and the like.

12. **Indemnification.** The Author agrees to indemnify JoVE and/or its successors and assigns from and against any and all claims, costs, and expenses, including attorney's fees, arising out of any breach of any warranty or other representations contained herein. The Author further agrees to indemnify and hold harmless JoVE from and against any and all claims, costs, and expenses, including attorney's fees, resulting from the breach by the Author of any representation or warranty contained herein or from allegations or instances of violation of intellectual property rights, damage to the Author's or the Author's institution's facilities, fraud, libel, defamation, research, equipment, experiments, property damage, personal injury, violations of institutional, laboratory, hospital, ethical, human and animal treatment, privacy or other rules, regulations, laws, procedures or guidelines, liabilities and other losses or damages related in any way to the submission of work to JoVE, making of videos by JoVE, or publication in JoVE or elsewhere by JoVE. The Author shall be responsible for, and shall hold JoVE harmless from, damages caused by lack of sterilization, lack of cleanliness or by contamination due to

the making of a video by JoVE its employees, agents or independent contractors. All sterilization, cleanliness or decontamination procedures shall be solely the responsibility of the Author and shall be undertaken at the Author's expense. All indemnifications provided herein shall include JoVE's attorney's fees and costs related to said losses or damages. Such indemnification and holding harmless shall include such losses or damages incurred by, or in connection with, acts or omissions of JoVE, its employees, agents or independent contractors.

13. **Fees.** To cover the cost incurred for publication, JoVE must receive payment before production and publication of the Materials. Payment is due in 21 days of invoice. Should the Materials not be published due to an editorial or production decision, these funds will be returned to the Author. Withdrawal by the Author of any submitted Materials after final peer review approval will result in a US\$1,200 fee to cover pre-production expenses incurred by JoVE. If payment is not received by the completion of filming, production and publication of the Materials will be suspended until payment is received.

14. **Transfer, Governing Law.** This Agreement may be assigned by JoVE and shall inure to the benefits of any of JoVE's successors and assignees. This Agreement shall be governed and construed by the internal laws of the Commonwealth of Massachusetts without giving effect to any conflict of law provision thereunder. This Agreement may be executed in counterparts, each of which shall be deemed an original, but all of which together shall be deemed to be one and the same agreement. A signed copy of this Agreement delivered by facsimile, e-mail or other means of electronic transmission shall be deemed to have the same legal effect as delivery of an original signed copy of this Agreement.

A signed copy of this document must be sent with all new submissions. Only one Agreement is required per submission.

### CORRESPONDING AUTHOR

Name:

Elise W.M. Hopman

Department:

Department of Psychology

Institution:

University of Wisconsin-Madison

Title:

Signature:



Date:

02/28/2019

Please submit a **signed** and **dated** copy of this license by one of the following three methods:

1. Upload an electronic version on the JoVE submission site
2. Fax the document to +1.866.381.2236
3. Mail the document to JoVE / Attn: JoVE Editorial / 1 Alewife Center #200 / Cambridge, MA 02140