



## The production effect and the generation effect improve memory in picture naming

Eirini Zormpa, Laurel E. Brehm, Renske S. Hoedemaker & Antje S. Meyer

To cite this article: Eirini Zormpa, Laurel E. Brehm, Renske S. Hoedemaker & Antje S. Meyer (2018): The production effect and the generation effect improve memory in picture naming, *Memory*, DOI: [10.1080/09658211.2018.1510966](https://doi.org/10.1080/09658211.2018.1510966)

To link to this article: <https://doi.org/10.1080/09658211.2018.1510966>



© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 24 Aug 2018.




[Submit your article to this journal](#)



[View Crossmark data](#)

## The production effect and the generation effect improve memory in picture naming

Eirini Zormpa <sup>a</sup>, Laurel E. Brehm<sup>a</sup>, Renske S. Hoedemaker<sup>a</sup> and Antje S. Meyer<sup>a,b</sup>

<sup>a</sup>Psychology of Language Department, Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands; <sup>b</sup>Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen, Nijmegen, the Netherlands

### ABSTRACT

The production effect (better memory for words read aloud than words read silently) and the picture superiority effect (better memory for pictures than words) both improve item memory in a picture naming task (Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory (Hove, England)*, 20(7), 655–666. doi:10.1080/09658211.2012.693510). Because picture naming requires coming up with an appropriate label, the generation effect (better memory for generated than read words) may contribute to the latter effect. In two forced-choice memory experiments, we tested the role of generation in a picture naming task on later recognition memory. In Experiment 1, participants named pictures silently or aloud with the correct name or an unreadable label superimposed. We observed a generation effect, a production effect, and an interaction between the two. In Experiment 2, unreliable labels were included to ensure full picture processing in all conditions. In this experiment, we observed a production and a generation effect but no interaction, implying the effects are dissociable. This research demonstrates the separable roles of generation and production in picture naming and their impact on memory. As such, it informs the link between memory and language production and has implications for memory asymmetries between language production and comprehension.

### ARTICLE HISTORY

Received 29 March 2018  
Accepted 7 August 2018

### KEYWORDS

Production effect; generation effect; picture superiority effect; picture naming; long-term memory

Over the decades, memory research has identified a set of encoding strategies that can enhance retention even after a single exposure to an item. One of these is simply producing a name aloud: reading a word aloud, as opposed to silently, improves recognition memory by 10 to 20%. This is known as the *production effect* (e.g., Conway & Gathercole, 1987; MacLeod & Bodner, 2017; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010). Although usually tested with words, this effect has also been reported in a picture naming task (Fawcett, Quinlan, & Taylor, 2012). Additionally, the production effect also arises across different manners of production, such as singing, whispering, mouthing, and typing (Forrin, Macleod, & Ozubko, 2012; Quinlan & Taylor, 2013). This effect has been most successfully explained under the *distinctiveness account*: The idea is that producing a name aloud adds an extra feature, that of the production record, to an item's memory trace (Conway & Gathercole, 1987; MacLeod et al., 2010; Ozubko & Macleod, 2010), which can be used heuristically at test to improve memory performance (Dodson & Schachter, 2001). Two further manipulations of encoding condition

that impact memory performance are, first, studying items as pictures versus as words (the *picture superiority effect*) and, second, generating labels for items versus reading existing labels (the *generation effect*). In this article, we examine how the production effect, the generation effect, and the picture superiority effect interrelate. We extend the claim that producing a picture name aloud enhances memory through speaking to propose that it also enhances memory through the active generation of a label. We test this claim in two recognition memory experiments where we dissociate the role of overt word production from the role of label generation in item memory.

In order to account for the effects of word production on picture naming, we first consider how pictures themselves influence memory compared to words. Picture stimuli tend to be remembered better than words, a finding known as the picture superiority effect (Paivio, Rogers, & Smythe, 1968). This effect is robust to a variety of manipulations, including modality changes between study and test. That is, memory for items studied as pictures is better than memory for items studied as words,

**CONTACT** Eirini Zormpa  eirini.zormpa@mpi.nl  Psychology of Language Department, Max Planck Institute for Psycholinguistics, P.O Box 310, 6500AH Nijmegen, the Netherlands

 Supplemental data for this article can be accessed <https://doi.org/10.1080/09658211.2018.1510966>.

© 2018 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

even when memory is tested with words (Borges, Stepanovsky, & Holt, 1977). Distinctiveness has also been successful in accounting for the picture superiority effect (Curran & Doyle, 2011; Pyc & Rawson, 2009). Pictures are thought to be remembered better than words because the visual representations evoked by pictures are more distinctive than those evoked by printed words (Nelson, Reed, & McEvoy, 1977). It has also been argued that pictures undergo more extensive conceptual processing than words (Weldon & Roediger, 1987). This suggests that the picture superiority effect and the production effect may have a common basis: Both effects rely on distinctiveness of memory traces as a result of rich encoding at study.

The question of how the picture superiority effect and the production effect jointly relate to item memory was tested by Fawcett et al. (2012). They hypothesised that if the picture superiority effect and the production effect both rely on distinctiveness, they should interact to give pictures named aloud an especially distinctive memory trace. To test this hypothesis, Fawcett and colleagues provided participants with picture and word stimuli at study and asked them to mouth the item names, producing speech movements but no overt speech, or to generate the item names using inner speech only. In a yes-no recognition task for the pictures and words, participants showed increased sensitivity for mouthed than for internally named items, showing a production effect, and increased sensitivity for pictures than for words, showing a picture superiority effect. The predicted interaction also occurred, such that mouthing benefited items more in the picture condition than in the word condition. Fawcett and colleagues interpreted this interaction as evidence that distinctiveness underlies both the production effect and the picture superiority effect.

In Fawcett and colleagues' study, the distinctiveness of pictures was attributed to their visual characteristics. However, note that when words and pictures are named - be it overtly or covertly - the processes mediating between the visual input and the linguistic representations of the names also differ. Picture naming is conceptually driven: Speakers need to identify the concept represented in the picture, select an appropriate lexical unit, and retrieve the corresponding phonological and articulatory commands (Dell & O'Seaghdha, 1992; Indefrey & Levelt, 2004; Levelt, Roelofs, & Meyer, 1999). By contrast, reading primarily involves the mapping of orthographic onto phonological and articulatory representations. This process typically entails the activation of semantic and conceptual representations, but it does not hinge on these processes, as evidenced by the fact that skilled readers can readily read non-words (Rosson, 1983; Theios & Muise, 1977). Thus, picture naming and word reading rely on different cognitive processes and these processes may affect item memory differentially.

The hypothesis that the differences between reading and picture naming in post-perceptual processes may contribute to differences in recognition memory is further

motivated by another observation from the memory literature, the generation effect (e.g., Bertsch, Pesta, Wiscott, & McDaniel, 2007; Slamecka & Graf, 1978). This is the finding that participants are better at remembering words that they generated themselves than words that they read. In a common version of the paradigm used to elicit this effect, participants either read antonym pairs (e.g., "hot - cold") or are presented with the first member of the pair and generate the antonym themselves (e.g., "hot - c\_\_\_"). Other ways to elicit the generation effect include the generation of words from synonyms (Slamecka & Graf, 1978), semantic associates (Begg, Snider, Foley, & Goddard, 1989), translations (Slamecka & Katsaiti, 1987), rhymes (Slamecka & Graf, 1978), and definitions (Forrin, Jonker, & MacLeod, 2014; MacLeod et al., 2010). In all of these cases, words that are self-generated are remembered better than words that are just read. Importantly, most of these generation rules require semantic processing. By contrast, no generation effect is observed for non-words generated using either rhyme or letter transposition rules (Nairne & Widner, 1987; Nairne, Pusey, & Widner, 1985), both of which are known to lead to a generation effect for words. This pattern suggests that the generation effect stems from the conceptual or semantic processing of the stimuli at study. The exact mechanism behind the generation effect has not been settled in the literature, but similar to the production effect and the picture superiority effect, distinctiveness has been proposed as a likely explanation (Gardiner & Hampton, 1985, 1988; Kinoshita, 1989).

Since picture naming relies more on conceptual processing than word reading does, this raises the question of whether part of the picture superiority effect observed in previous work may be attributable to a generation effect, arising during the conceptual and linguistic encoding rather than the visual encoding of the stimuli (for similar proposals see Smith & Magee, 1980; Weldon & Roediger, 1987). Therefore, it is crucial to dissociate which processes contribute to the memory benefit for named pictures.

Beyond its impact on fundamental memory research, discovering the role of the generation effect and the production effect in picture naming also provides valuable insights for psycholinguistics. Recent work has reported memory asymmetries as a function of speaking versus listening. More specifically, language production<sup>1</sup> seems to elicit a memory benefit, such that speakers have superior item memory compared to listeners for items referred to in a conversational context (Hoedemaker, Ernst, Meyer, & Belke, 2017; McKinley, Brown-Schmidt, & Benjamin, 2017; Yoon, Benjamin, & Brown-Schmidt, 2016). Disassociating the roles of the generation and production effects, both of which relate to mechanisms of language production, allows us to characterise the asymmetries between language production and comprehension and, more broadly, provides insight on how language and memory intertwine.

## Current study

The main goal of the present study was to examine whether the generation of picture names would improve recognition memory for the pictures via a more distinctive conceptual and linguistic representation of the stimuli. To ensure that generation rather than visual distinctiveness was the factor influencing performance, participants were presented with pictures in all conditions of our study. In one condition (picture + word condition), picture names were provided as labels superimposed on the pictures, whereas in the other condition (picture-only condition), labels were replaced by random patterns of equal visual complexity requiring participants to generate the names themselves (see also Weldon & Roediger, 1987). This allowed us to compare a picture-only to a picture + word condition rather than comparing a picture to a word condition, making the two stimulus types well matched with respect to visual distinctiveness. The prediction was that the presence of the correct labels should greatly facilitate the generation of the object names in the picture + word condition. One may think of the correct labels as identity primes for the picture names. If the conceptually driven generation of object names facilitates later recognition of the objects, performance in a recognition test should be better in the picture-only than in the picture + word condition.

Following Fawcett et al. (2012), we also manipulated whether the pictures were overtly named or not, aiming to elicit a production effect. In the silent condition, participants produced the object names in inner speech, and in the aloud condition, they produced the object names aloud. In the latter condition, we elected to use overt rather than mouthed speech, unlike Fawcett et al. (2012), because we wanted to measure the participants' response latencies. This meant that the effects of the engagement of articulatory gestures and auditory feedback were confounded, but separating them was not of interest in the current study. For evidence that audition improves memory see MacLeod (2011) and Forrin and MacLeod (2017).

The study design followed Fawcett and colleagues' Experiment 3. The four conditions created by crossing stimulus and response type were intermixed throughout the experiment, with visual cues (coloured frames around the pictures) indicating whether the picture was to be named aloud or silently. Deviating from the earlier study, items were presented in a delayed naming paradigm (discussed in the Method section of Experiment 1) in order to encourage participants to generate object names in both the aloud and the silent conditions. After the naming task, participants completed a 20-minute filler test – a computerised version of Raven's Advanced Progressive Matrices (Raven, Raven, & Court, 1998). The final experimental task was a self-paced Yes/No recognition memory test.

In Yes/No recognition tests, participants' tendency to respond positively or negatively can influence hit rates.

This confound is avoided by computing each person's sensitivity and bias in line with signal detection theory (Macmillan & Creelman, 2005). Here we use an application of signal detection theory to mixed-effect models using log-odds ratios (DeCarlo, 1998; Wright, Horry, & Skagerberg, 2009). The benefit of this approach is that it accounts for participant and item variability. As with any other signal detection analysis, the primary dependent measure was sensitivity. We predicted higher sensitivity for items in the picture-only condition than in the picture + word condition, reflecting the generation effect. We also predicted higher sensitivity in the aloud condition than in the silent condition, reflecting the production effect. Assuming that the generation and production effects stem from different aspects of the language production system, we predicted no interaction between stimulus type and response type. This is consistent with the findings of Experiment 1 of Forrin et al. (2014), in which the contributions of the generation and production effects on recognition memory for words were found to be independent.

## Experiment 1

This experiment was designed to test the role of the generation effect and the production effect in recognition memory for items studied in a picture naming task. It consisted of three tasks run in one session of approximately one hour. Participants first completed a picture naming task (the study phase) where items varied in their stimulus type (whether the name was primed by the label or not) and in the type of response they elicited (aloud or silent). Participants then completed a filler task, and finally a forced-choice recognition memory task (the test phase).

## Method

### Participants

Forty-three native Dutch speakers (nine male, age range: 18–30 years) participated in the experiment. They were recruited from the Max Planck Institute for Psycholinguistics participant database and received €10 each. All participants had normal colour vision. One participant was excluded due to experimenter error and another due to slow response times, leaving 41 participants in the analysis. A power calculation in G\*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that with 32 items per condition this sample was sufficient to detect effects of at least Cohen's  $f = 0.49$  (a medium to large effect size) with 95% power<sup>2</sup>. Ethical approval to conduct the study was given by the Ethics Board of the Social Sciences Faculty of Radboud University.

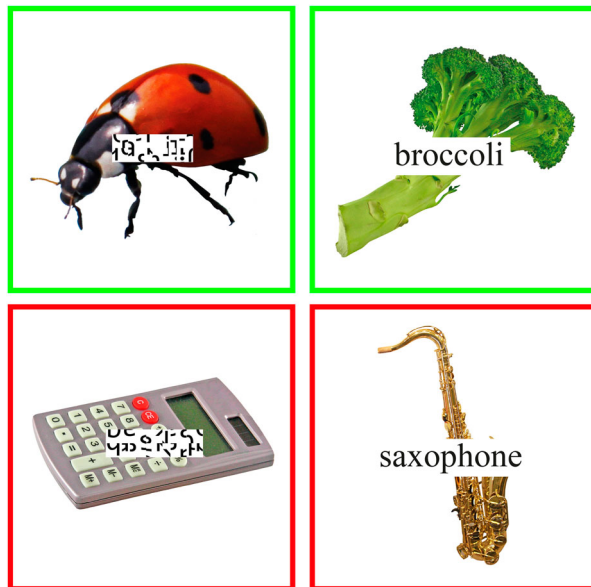
### Materials and design

The stimuli were 256 pictures selected from the BOSS photo database, presented in 400 × 400 pixel resolution (Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010;

Brodeur, Guerard, & Bouras, 2014). Because the norming data from the BOSS database refer to the English names of the objects, name agreement scores were obtained from eight Dutch native speakers who did not participate in the experiments. Frequency scores for Dutch were obtained from the SUBTLEX-NL corpus (Keuleers, Brysbaert, & New, 2010). The 256 items were divided into two sets (A and B), which were matched on name agreement ( $M_A = .81$ ,  $M_B = .71$ ;  $t(254) = -.18$ ,  $p = .40$ ),  $\log_{10}$  word frequency ( $M_A = .216$ ,  $M_B = 2.20$ ,  $t(254) = -.33$ ,  $p = .75$ ), familiarity ( $M_A = 4.35$ ,  $M_B = 4.38$ ,  $t(254) = -.58$ ,  $p = .26$ ), visual complexity ( $M_A = 2.36$ ,  $M_B = 2.34$ ,  $t(254) = .33$ ,  $p = .70$ ), manipulability ( $M_A = 2.79$ ,  $M_B = 2.88$ ,  $t(254) = -.84$ ,  $p = .19$ ), and length (in letters) ( $M_A = 6.70$ ,  $M_B = 7.03$ ,  $t(254) = -.93$ ,  $p = .73$ ).

Eight lists were constructed using the two picture sets. In four lists, pictures from set A were used as targets: They were named at study and presented again at test as old items, and set B pictures were presented as foils at test only. In the four remaining lists, set B pictures were used as targets and set A pictures as foils.

At *study*, participants saw each of the 128 items in one of the four naming conditions created by crossing stimulus type and response type (picture-only aloud, picture-only silent, picture + word aloud, and picture + word silent). There were 32 items per condition; assignment of items to conditions was counterbalanced across lists. In the picture + word condition, the names of the objects were presented as printed labels in 20 pt Arial



**Figure 1.** Examples of the stimuli used for each naming condition. Top left: picture-only aloud, top right: picture + word aloud, bottom left: picture-only silent, bottom right: picture + word silent. Green frames signalled participants to produce items aloud and red frames signalled them to produce items silently. In the picture + word condition, participants were instructed to use the provided labels to name the pictures. In the picture-only condition, they were instructed to come up with an appropriate word on their own.

font superimposed upon the pictures. In the picture-only condition, the printed labels were replaced by random patterns of matched visual complexity. These patterns were created from non-words which were formed by combining the ending of a target word to the beginning of a different target word. Each letter of the non-words was divided into segments ( $3 \times 3$  grid), which were then assigned random locations and orientations. To indicate the silent versus aloud conditions, coloured frames were presented around the items: green (RGB: 0, 255, 0) for aloud production and red (RGB: 255, 0, 0) for silent production (see Figure 1 for stimuli examples).

At *test*, all 256 pictures were included, with probe type (target versus foil) manipulated between items. The 128 studied pictures served as targets and the 128 non-studied pictures as foils. Target pictures appeared as they had at study, i.e., with or without superimposed printed names and with a green or a red frame. Foil pictures also appeared with a green or red frame around them and a superimposed readable or unreadable label, as if they had belonged to a naming condition. This enabled us to calculate separate false alarm rates for each condition (see also Fawcett et al., 2012; Forrin, Groot, & MacLeod, 2016).

#### Procedure

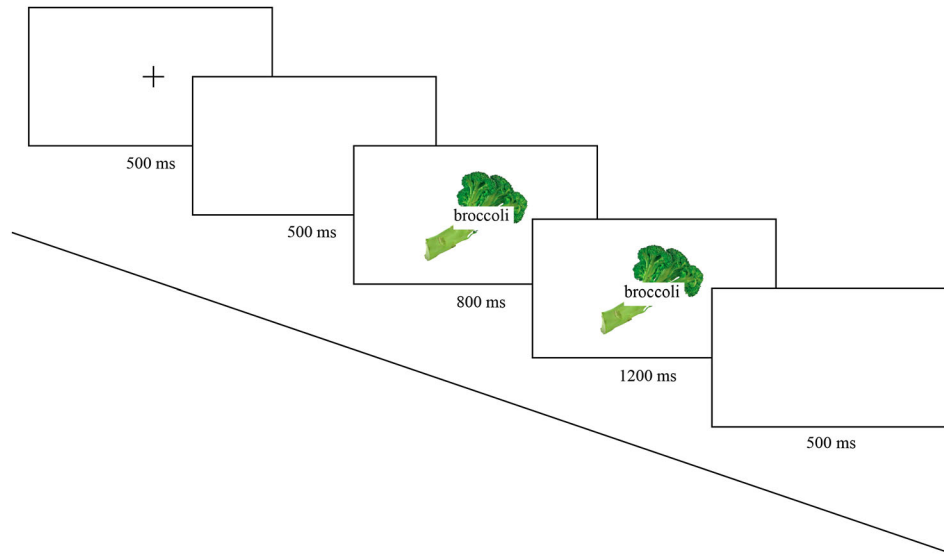
The experiment was controlled by Presentation (Neurobehavioral Systems) and displayed on a 24'' monitor (1920  $\times$  1080 pixel resolution).

**Picture naming.** Trials started with a fixation cross presented in the middle of the screen for 500 ms, followed by a blank screen for 500 ms. Next, a picture appeared in the middle of the screen. After 800 ms a green or red frame appeared around the picture, and both picture and frame remained on screen for 1200 ms. A blank screen appeared for 500 ms, serving as an inter-trial interval (see Figure 2 for an example of the sequence of events in this task). Participants were instructed to start preparing the picture name as soon as they saw the picture, but to produce it only after seeing the frame. As described in the Introduction, a delayed naming paradigm was used to make sure that the participants generated the picture names both in the silent and the aloud condition.

The experiment was preceded by a practice phase of 12 trials, in which 12 pictures that did not appear in the main experiment were used. Three pictures appeared in each of the naming conditions. The practice trials had the same structure as experimental trials with the exception that the reaction time for each aloud trial was displayed on the screen before the next trial began to encourage participants to respond as fast as possible.

**Filler task.** The filler task was a computerised version of Raven's Advanced Progressive Matrices (Raven et al.,





**Figure 2.** Example of the sequence of events during a trial in a picture + word aloud condition.

1998). It consisted of 36 multiple-choice problems that become progressively more difficult. In each trial, participants saw a pattern which needed to be completed and eight possible solutions. Participants had 20 min to solve as many problems as possible.

**Recognition memory.** Recognition memory was tested in a self-paced Yes/No memory task. Participants saw one picture at a time and were instructed to press the left arrow key if they had named the picture during study and the right arrow key if they had not done so. As a reminder, “JA” (yes) was printed at the bottom left corner of the screen and “NEE” (no) at the bottom right. Trial order was randomised within Presentation.

### Analysis

Memory performance was the primary dependent variable in this study. Naming data were used only to exclude trials.

**Naming.** The naming task was designed to encourage participants to prepare to name all the pictures they saw, even those in the silent naming condition. To test the effectiveness of this design we ran a norming experiment to estimate the average naming times for our stimuli and contrast them with the average naming times in the main experiment. In the norming experiment, eight Dutch native speakers who did not participate in the main experiment named aloud all the pictures. Their responses were annotated in Praat (Boersma & Weenink, 2016) by the first author and their naming latencies were extracted. In the picture-only condition, norming participants started speaking 1000 ms ( $SD$  456 ms) after picture onset. In contrast, participants in the main study started speaking 477 ms ( $SD$  185) after the frame (the response cue) was presented. In the picture + word condition, norming participants started speaking

705 ms ( $SD$  367 ms) after picture onset and main study participants started speaking 402 ms ( $SD$  84 ms) after the frame was presented. This demonstrates that in both overt conditions, participants began planning as soon as they saw the image, allowing them to respond faster than the norming participants when given the cue to produce a response. Therefore, we concluded that our participants covertly named the pictures in the silent conditions. One individual with unusually slow naming times ( $M = 572$  ms) was excluded, as mentioned in the Participants section.

Also excluded from analysis were trials in which participants named pictures incorrectly or did not follow the naming instruction (1.03% of all trials). In addition, aloud trials were excluded in which participants repeated a word from a previous trial. For example, when a participant named a cigarette correctly, but then also named a piece of chalk *cigarette*, both correct and incorrect trials were excluded (0.2% of all trials) because repeated words (i.e., *cigarette*) would likely have a memory advantage over words said once. For similar reasons, when an incorrect word was used during naming and this incorrect word was the name of a foil used in the memory test, the memory trials for both the target and the foil were excluded (0.1% of all trials). For example, if a participant named a dolphin *shark* and then saw a shark during the memory task, the naming trial and the memory trials for *shark* and *dolphin* were excluded.

**Memory.** Responses were modelled as a function of probe type (target/foil), stimulus type (picture-only/picture + word), response type (aloud/silent), and their interactions (Wright et al., 2009). This provides measures of bias (how likely participants were to say Yes to any item) and sensitivity (how likely participants were to say Yes to an old item), and the way that these measures

were modulated by the predictors; as such the analysis is conceptually similar to the standard signal detection analysis and is consistent with current best practices (e.g., Fawcett & Ozubko, 2016; Fraundorf, Watson, & Benjamin, 2010; Jacobs, Dell, Benjamin, & Bannard, 2016; Yoon et al., 2016).

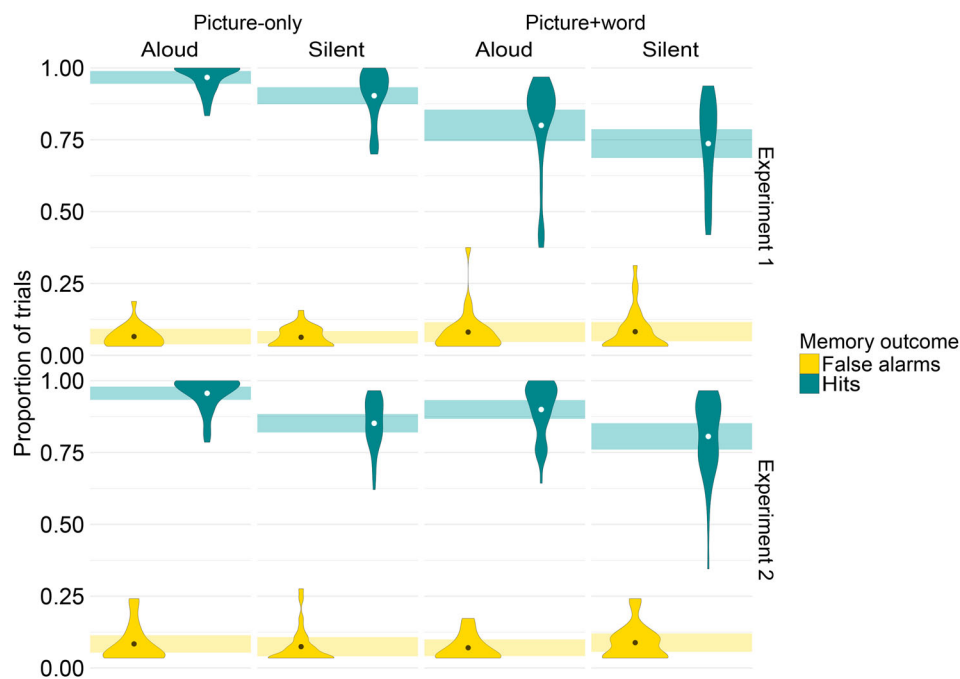
Memory performance was analysed using a logistic mixed effects model that was run using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2015) in R (R Development Core Team, 2014) with the optimiser *bobyqa* (Powell, 2009). All coefficients represent changes in log-odds of Yes versus No responses as a function of the predictor. No responses were coded as 0 and Yes responses as 1. All fixed effects were sum-to-zero coded with foils, picture + word and silent conditions coded to  $-.5$  and targets, picture-only, and aloud conditions to  $+.5$ . The model with the maximal random effects structure included random intercepts for participants and items, random slopes for the effects of stimulus type, response type and their interaction on items, and random slopes for the effects of probe type, stimulus type, response type and their interactions on participants. This model failed to converge, so the interaction terms for the by-subject and by-item random slopes were removed. In this model, the random slopes for stimulus type and response type effects on items were highly correlated with the random intercept for items (1.00 in both cases), indicating overfitting, and so both random slopes were excluded. Furthermore, the by-item random slope for response type was removed because it explained very little variation (0.05). The final model included random intercepts for participants and

items, as well as random slopes by subject of probe type and stimulus type. Reported  $p$ -values were obtained using likelihood ratio tests comparing the full model to a model with the same random effects structure but without the fixed effect in question, which has been shown to be more conservative than Wald  $z$  tests for small samples (Agresti, 2007). Reported 95% confidence intervals were obtained using the *confint* function.

## Results

Accuracy in the memory task was high (90%). As Figure 3 shows, the hit rates were higher in the picture-only than in the picture + word condition, and they were higher in the aloud than in the silent naming condition. In contrast to the hit rates, the false alarm rates were minimally affected by the experimental conditions. This meant that, in a signal detection analysis, parallel results were obtained for sensitivity and bias. We focus on the sensitivity results below as they are of primary interest.

The full logistic regression model is shown in Table 1<sup>3</sup>. Sensitivity was evaluated using the interactions between probe type (target versus foil) and the other predictors (stimulus type and response type). Consistent with the generation effect, probe type interacted with stimulus type such that participants were more likely to correctly recognise studied items in the picture-only than in the picture + word condition ( $\beta = 2.01$ ,  $z = 10.86$ , 95% CIs [1.63, 2.40],  $p < .001$ ). Consistent with the production effect, probe type interacted with response type such



**Figure 3.** Proportions of false alarms and hits for each naming condition (picture-only aloud, picture-only silent, picture + word aloud, picture + word silent) for Experiments 1 and 2 (panels). False alarms are calculated from the foils in the memory task and hits from the targets. Dots represent means by condition. Rectangles represent normalised within-participant 95% confidence intervals by condition.

**Table 1.** Experiment 1. Mixed-effects logistic regression of the log-odds of Yes responses for all 256 trials. Stimulus type was picture-only or picture + word, Response type was aloud or silent, and Probe type was target or foil. See text for *p* values.

		Estimate	SE	Wald z			Variance	SD
<i>Bias</i>	<i>Fixed</i>				<i>Random</i>			
	Intercept	−0.52	0.11	−4.79	<i>Participant</i>	Intercept	0.32	0.56
	Stimulus type	0.84	0.11	7.57		Probe	1.39	1.18
	Response type	0.31	0.08	3.62		Stimulus	0.13	0.36
	Stimulus:Response	0.45	0.17	2.69	<i>Item</i>	Intercept	0.40	0.63
<i>Sensitivity</i>	Probe type	5.84	0.22	27.14				
	Probe:Stimulus	2.01	0.18	10.86				
	Probe:Response	1.07	0.17	6.36				
	Probe:Stimulus:Response	0.74	0.34	2.21				

that participants were more likely to correctly recognise studied items in the aloud than in the silent condition ( $\beta = 1.07$ ,  $z = 6.36$ , 95% CIs [0.73, 1.42],  $p < .001$ ). An additional interaction between probe type, stimulus type, and response type was also observed, such that studied items in the picture-only condition benefited from production more than those in the picture + word condition ( $\beta = 0.74$ ,  $z = 2.21$ , 95% CIs [0.06, 1.44],  $p < .05$ ).

## Discussion

This experiment yielded two main findings. First, response type impacted memory performance: Saying the picture names aloud during study facilitated picture recognition compared to naming in inner speech. This suggests that better recognition memory stemmed from overt articulation during study. This pattern is consistent with numerous findings in the literature demonstrating a production effect in recognition memory performance (Bodner, Jamieson, Cormack, McDonald, & Bernstein, 2016; Forrin et al., 2012; MacLeod et al., 2010; MacLeod & Bodner, 2017; Mama & Lcht, 2016).

Second, participants' recognition performance was impacted by stimulus type, such that it was better when participants had to generate the object names than when the object names were provided as identity primes in the form of written labels superimposed on the pictures. This result may be seen as a generation effect for picture names, similar to generation seen in other paradigms, e.g., when participants read or generate antonyms (Slamecka & Graf, 1978). Naming pictures may be well-practiced, but it is not instantaneous, as it requires conceptual and linguistic processing (Indefrey & Levelt, 2004; Levelt et al., 1999). Although conceptual processing is not necessary to elicit a generation effect, it has been repeatedly shown to be sufficient to do so (Nairne & Widner, 1987; Weldon & Roediger, 1987). This finding is therefore consistent with the literature at large.

However, the interpretation of this finding is complicated by a property of the experimental design. In Experiment 1, all labels correctly indicated the object names. Thus, instead of using the labels to facilitate naming of the pictures, as we assumed, the participants may have entirely relied on the labels and ignored the pictures. This would mean that, despite pictures being present on all trials, the experiment may not have assessed the effect of aided (by

identity prime) versus unaided picture naming, but instead may have assessed the effect of word reading versus picture naming. This would entail that the pictures may have been processed less attentively in the picture + word than in the picture-only condition. Therefore, it is possible that this difference in picture processing and not the act of generating the object names can account for the better memory that was found in the picture-only than in the picture + word condition. In other words, the picture superiority effect and not the generation effect may be responsible for the pattern in our data. Experiment 2 was conducted to assess this concern.

We also observed an unanticipated experimental finding: Stimulus type and response type had an overadditive effect such that overt production was more beneficial in the picture-only than in the picture + word condition. This effect is reminiscent of the pattern reported by Fawcett et al. (2012), who found an overadditive effect of stimulus type and response type that they attributed to an interaction between the picture superiority effect and the production effect. If the picture superiority effect did in fact affect our results, then this overadditive effect might also be explained by an interaction between the picture superiority effect and the production effect.

## Experiment 2

The goal of Experiment 2 was to minimise the impact of the picture superiority effect by making the picture labels occasionally unreliable. To encourage full processing of the images even in the picture + word conditions, we reasigned 10% of the trials in the naming task to be catch trials. On these trials there was a mismatch between the picture and the label, meaning that participants could no longer be certain that the word they read was a suitable name for the picture they saw. This manipulation was designed to force participants to always look at the pictures in order to evaluate the reliability of the labels with which they were presented, thus making picture processing necessary in all conditions. As a result of this longer visual processing, we expected sensitivity in the picture + word conditions to increase relative to Experiment 1. However, we still predicted effects on sensitivity based upon stimulus type, with higher sensitivity in the picture-



only than in the picture + word condition, as well as effects based upon response type, with higher sensitivity in the aloud condition than in the silent condition. Furthermore, if the interaction reported in Experiment 1 was driven by the picture superiority effect, and if the picture superiority effect was now eliminated, we should now see additive, rather than overadditive effects of stimulus type and response type.

## Method

### Participants

Forty-two participants (9 males, age range: 18–30) who had not participated in Experiment 1 were recruited from the Max Planck Institute participant database and received €10 for their participation. One participant was excluded because of low performance in the memory test (accuracy < 25%), leaving 41 participants in the analysis.

### Materials and design

The same materials and design were used as in Experiment 1 with the exception that 12 items in each picture set were reassigned to be catch trials. This means that there were 116 critical trials and 12 catch trials in the naming task. The labels presented on catch trials were semantically and phonologically unrelated to the pictures. Target picture names and distractor labels were of comparable length ( $M_T=7.5$ ,  $M_D=6.9$ ),  $\log_{10}$  word frequency ( $M_T=2.14$ ,  $M_D=2.27$ ), familiarity ( $M_T=4.27$ ,  $M_D=4.35$ ), and manipulability ( $M_T=2.8$ ,  $M_D=3.0$ ). Three catch items appeared per naming condition.

Catch trials were not included in the memory task. This means that 12 items from each picture set were excluded from the memory test, leaving 232 trials. To ensure that this length difference would not affect our results, data from the first 232 trials in Experiment 1 were reanalysed. Results from the full and shortened datasets did not differ.

### Procedure

The procedure was identical to Experiment 1, but for the inclusion of catch trials in the practice and experimental trials.

### Analysis

The analysis was identical to Experiment 1.

**Naming.** Naming latencies were calculated to ensure that participants prepared to name the pictures as soon as they were shown. These were compared to the naming latencies of the norming experiment presented in Experiment 1, where participants were instructed to name the pictures as soon as they saw them. In the picture-only condition, norming participants started speaking approximately 1000 ms after picture onset and main study participants started speaking 499 ms ( $SD$  188 ms) after the frame (the response cue) was presented. In the picture + word condition, norming participants started speaking

approximately 705 ms after picture onset and main study participants started speaking 454 ms ( $SD$  137 ms) after the frame was presented. This implies that, as instructed, participants began planning the object names before the response cue was given.

Trials in which participants gave an incorrect response or did not follow the naming instruction were excluded (1.3% of all data). Trials in which participants used the same word to name two objects were excluded (0.2% of all trials), as were trials in which pictures were named incorrectly with the name of a foil (0.1% of all trials).

**Memory.** A logistic mixed effects model was run using the lme4 package (Bates et al., 2015) in R (R Development Core Team, 2014) with the optimiser *bobyqa* (Powell, 2009). The same contrasts were used as in Experiment 1. As in Experiment 1, the maximal model failed to converge so the interaction terms for the by-subject and by-item random slopes were removed. Although this model converged, the random slopes for stimulus type and response type effects on participants were highly correlated with the random intercept for participants (−0.96 and 0.99 respectively), indicating overfitting, and were therefore both excluded from the final model. This model included random intercepts for participants and items and random slopes for the effects of stimulus and response type on items, plus a random slope for the effect of probe type on participants. As in Experiment 1,  $p$ -values were obtained using likelihood ratio tests and 95% confidence intervals were obtained using the *confint* function.

## Results

Accuracy in the memory task was overall very high (91%). Hit rates were higher in the picture-only and the aloud conditions than in the picture + word and the silent conditions (Figure 3). As before, the false alarm rates were unaffected by condition, meaning that the bias and sensitivity measures follow identical patterns. We again discuss sensitivity only for brevity.

The full logistic regression model is shown in Table 2. Sensitivity was measured using the interactions between probe type and the other two predictors (stimulus type and response type). As in Experiment 1 there was an interaction between probe type and stimulus type ( $\beta = 0.64$ ,  $z = 3.10$ , 95% CIs [0.24, 1.05],  $p < .001$ ), meaning that participants were significantly more likely to correctly recognise items in the picture-only than in the picture + word condition. There was also an interaction between probe type and response type ( $\beta = 1.49$ ,  $z = 7.13$ , 95% CIs [1.09, 1.91],  $p < .001$ ), meaning that participants were more likely to correctly recognise items in the aloud than in the silent condition. The interaction between probe type, stimulus type, and response type was not significant ( $\beta = -0.09$ ,  $z = -0.24$ , 95% CIs [−0.81, 0.64],  $p > 0.9$ ).

**Table 2.** Experiment 2. Mixed-effects logistic regression of the log-odds of Yes responses for all 232 trials. Stimulus type was picture-only or picture + word, Response type was aloud or silent, and Probe type was target or foil. See text for *p* values.

		Estimate	SE	Wald z			Variance	SD
<i>Bias</i>	<i>Fixed</i>				<i>Random</i>			
	Intercept	-0.47	0.12	-4.00	<i>Participant</i>	Intercept	0.39	0.62
	Stimulus type	0.34	0.09	3.77		Probe	1.03	1.01
	Response type	0.49	0.09	5.45	<i>Item</i>	Intercept	0.38	0.61
	Stimulus:Response	0.59	0.17	3.37		Stimulus	0.14	0.38
<i>Sensitivity</i>	Probe type	5.86	0.20	28.85		Response	0.09	0.30
	Probe:Stimulus	0.64	0.21	3.10				
	Probe:Response	1.49	0.21	7.13				
	Probe:Stimulus:Response	-0.88	0.40	-0.24				

## Discussion

Experiment 2 differed from Experiment 1 only in the inclusion of catch trials with incongruent labels in the study phase. The purpose of the catch trials was to make it impossible for participants to exclusively rely on the labels for naming and thus force them to process the pictures in both the picture-only and the picture + word condition. This should make the extent to which generation takes place the main difference between stimulus types. The results largely replicate those of Experiment 1. We observed an effect of stimulus type on sensitivity, with better recognition memory when the name of the picture needed to be generated without rather than with the support of a matching label. This supports the existence of a generation effect in picture naming. We also observed an effect of response type on sensitivity, with better recognition memory in the aloud than in the silent condition, consistent with the existence of a production effect in picture naming. These effects were additive: the interaction reported in Experiment 1 was not replicated.

## General discussion

The goal of the present study was to test the hypothesis that producing picture names aloud enhances memory not only by the act of speaking but also by the process of actively generating labels, such that the generation of picture names improves memory above and beyond the benefit coming from the distinctiveness of the visual features of pictures. In two experiments, we manipulated stimulus type and response type, such that participants generated or read picture names aloud or silently. Our design had two primary methodological differences from Fawcett et al. (2012). First, we used a picture + word condition instead of a word condition in an attempt to neutralise the visual distinctiveness of pictures and assess whether the generation of labels still improved memory. Second, we used delayed naming to ensure that participants were engaging fully in the task of generating names even when they were not expected to make an overt response. In both experiments, we showed that both stimulus type (picture-only/picture + word) and response type (aloud/silent) affected memory performance. This was evidenced by higher sensitivity scores for the picture-only compared to the picture + word conditions, consistent with the generation effect,

and higher sensitivity scores for the aloud than for silent conditions, consistent with the production effect.

Experiment 1 also revealed an overadditive effect of stimulus type and response type on sensitivity such that responding aloud benefited items more in the picture-only than in the picture + word condition. Our aim in Experiment 1 was to test whether a generation effect could be captured in picture naming, and therefore we included pictures in all conditions. However, it is possible that the reliability of the labels in the picture + word condition led attention to be directed away from the pictures, making the visual features of the picture more prominent in the picture-only than in the picture + word condition and serving to enhance memory. We addressed this concern in Experiment 2 with the inclusion of unreliable labels, i.e., labels other than the names of the pictures they accompanied. The fact that participants could no longer exclusively rely on the labels meant that they always had to attend to the pictures. We expected this to reinforce visual processing of the pictures and thus improve performance in the picture + word condition. Hit rates in the picture + word condition were higher in Experiment 2 than in Experiment 1, supporting this claim. The interaction between stimulus type and response type disappeared in Experiment 2, consistent with our characterisation of this interaction as resulting from differential picture processing in the picture + word condition across experiments.

To ensure the validity of the comparison of sensitivity across Experiments 1 and 2, we ran a logistic regression with experiment as a fixed effect on the pooled data from both experiments ( $N = 82$ ). Experiment 1 was coded as  $-0.5$  and Experiment 2 as  $+0.5$ ; the other effects were coded as previously. Results of this analysis are shown in Table 3. Importantly, the interaction between probe type and stimulus type, which captures sensitivity towards the two different types of stimuli, changed significantly as a function of experiment. We interpret this as a result of the increase in hit rates in the picture + word condition in Experiment 2, evident in Figure 3. Despite this change, sensitivity remained greater for items in the picture-only condition than in the picture + word condition both in Experiment 2 and in the pooled data from Experiments 1 and 2. This provides strong evidence that generation in a picture naming task enhances recognition memory. That is, the process of generating a label for a picture at encoding enhances memory at a later test.

**Table 3.** Comparison of Experiments 1 and 2. Mixed-effects logistic regression of the log-odds of Yes responses for first 232 trials. Stimulus type was picture-only or picture + word, Response type was aloud or silent, Probe type was target or foil. See text for *p* values.

		Estimate	SE	Wald z			Variance	SD
	<i>Fixed</i>					<i>Random</i>		
<i>Bias</i>	Intercept	-0.48	0.09	-5.43	<i>Participant</i>	Intercept	0.37	0.61
	Stimulus type	0.56	0.08	7.31		Probe	1.21	1.10
	Response type	0.39	0.07	5.38		Stimulus	0.06	0.24
	Stimulus:Response	0.49	0.12	3.99	Response	0.39	0.20	
	Experiment	0.01	0.16	0.06	Exper.	0.06	0.25	
	Experiment:Stimulus	-0.39	0.17	-2.86	<i>Item</i>	Intercept	0.38	0.63
	Experiment:Response	0.21	0.13	1.57		Stimulus	0.15	0.38
Experiment:Stimulus:Response	0.17	0.25	0.48	Response		0.06	0.25	
<i>Sensitivity</i>	Probe type	5.90	0.15	39.38	Exper.	0.05	0.23	
	Probe:Stimulus	1.39	0.16	8.90				
	Probe:Response	1.45	0.15	9.41				
	Probe:Experiment	-0.03	0.29	-0.12				
	Probe:Stimulus:Response	0.40	0.26	1.54				
	Probe:Experiment:Stimulus	-1.55	0.27	-5.64				
	Probe:Experiment:Response	0.12	0.26	0.48				
Probe:Exper:Stimulus:Response	-0.76	0.49	-1.55					

In contrast to the effects of stimulus type, which varied between experiments, response type disclosed consistent patterns across both experiments, clearly demonstrating that overt production of picture names greatly improves recognition memory for the pictures compared to production in inner speech. This production effect for picture names was first reported by Fawcett et al. (2012) in three experiments that compared mouthing to inner speech. The two experiments reported here compared overt speech to inner speech and again found a production effect (see also Richler, Palmeri, & Gauthier, 2013). The results of these studies suggest that the production effect for pictures, like the production effect for words, is robust across different means of production.

This study is the first to report a simultaneous role for the production effect and the generation effect in picture naming. This converges with evidence provided by Experiment 1 of Forrin et al. (2014), in which the production and the generation effect independently improved recognition memory for words. It also converges with a larger body of earlier work. For instance, Fawcett et al. (2012) had previously reported a production effect in picture naming by showing that pictures that were named overtly were remembered better than pictures that were named in inner speech. They further showed that pictures are remembered better than words, although it is not possible to discriminate between the picture superiority effect and the generation effect from this comparison. Furthermore, Weldon and Roediger (1987) in their Experiment 2 used a picture-only versus picture + word contrast and found better memory in the picture-only condition, indicative of a generation effect. However, in that experiment, all words were produced in inner speech. The present study combined the methodology of previous work to find that the production effect and the generation effect independently improve memory in picture naming.

The findings of this study have broader implications regarding the generation effect and the way it relates to the picture superiority effect. Regarding the generation effect, we have extended its boundaries to include the naming of intact pictures. Previously, Kinjo and Snodgrass

(2000) reported a generation effect for pictures by comparing the naming of fragmented pictures to the naming of intact ones. Here we have shown that the naming of intact pictures, which also requires generation in the form of conceptual or linguistic processing, is able to give similar memory benefits. This corroborates the findings reported by Weldon and Roediger (1987) in their Experiment 2 and underscores the role of picture naming in the generation effect in addition to its role in the picture superiority effect. We have also shown that the generation affect and the picture superiority effect are very closely linked in picture naming tasks, which has implications for the methodology used to study the picture superiority effect. That is, studies primarily interested in the visual distinctiveness of pictures need to carefully control for the differences in the conceptual and linguistic processing required for pictures versus words.

In the Introduction, we noted that the production effect, the picture superiority effect, and the generation effect are often all ascribed to enhanced distinctiveness of the memory representations of the target items. If a common processing principle underlies all effects, one might expect the effects to interact; and indeed this prediction is often borne out (e.g., Fawcett et al., 2012). In Experiment 1 of the present study, an interaction of stimulus type and response type was seen, but this interaction was not replicated in Experiment 2. We interpret this pattern as indicating that the stimulus type effect of Experiment 1 was a compound of the generation effect and the picture superiority effect. When the picture superiority component was eliminated in Experiment 2, the interaction disappeared. Consequently, it appears that in our study, the production effect may have interacted with the picture superiority effect, whereas the production effect and generation effect did not interact. This pattern may be seen to challenge the view that the latter two effects originated from the involvement of a shared mechanism. Alternatively, similar processing mechanisms may be implicated but applied to distinct representations. Thus, the picture superiority effect (not directly assessed in the present study) may be due to an enhancement of the distinctiveness of the visual representation of

the target (Nelson et al., 1977); the production effect may arise due to increased distinctiveness of the phonetic representation of the target name and the associated articulatory and motor commands (Forrin et al., 2012); and the generation effect in picture naming may arise from increased distinctiveness of the conceptual representation associated with the picture name. Note that the generation effect in picture naming discussed here for the first time is different from the “classic” generation effect for words. The participants did not generate the stimulus (the picture) as they would in the classic case, where a probe (“hot”) elicits a target (“cold”); instead they generated conceptual and lexical representations that were associated with the targets. Our results show that these processes enhance memory for the target pictures. Exactly how this effect arose needs to be elucidated in further work.

Our study also has implications for psycholinguistic research on the interface of language and memory and, more specifically, for the finding that speakers remember their own utterances better than their listeners do (Hoedemaker et al., 2017; Knutsen & Le Bigot, 2014; McKinley et al., 2017; Yoon et al., 2016). Although in the present study we only tested people individually, we report a similar finding: When participants acted as speakers (i.e., when they generated words or when they spoke aloud, their memory improved relative to when they did not engage in these activities). Thus, the current work provides a useful starting point for research into memory asymmetries between speakers and listeners in conversation.

## Conclusion

Generating a word from a cue improves memory for that item relative to reading. In two experiments, we demonstrate that a similar memory benefit arises from a picture naming task. Producing the name of the picture aloud improves memory (the production effect), as does generating a label for the picture (the generation effect). This demonstrates the interplay between language and memory and has implications for both memory and psycholinguistic research.

## Notes

- Note that the production effect refers to the comparison of overt versus inner speech. Language production refers to the entire process of producing language starting with the conceptual processing of what is being communicated and culminating in articulation.
- In G\*Power, the test family selected was “F tests”, the statistical test was “ANOVA: Repeated measures, between factors”, and the type of power analysis was “Sensitivity: Compute required effect size – given  $\alpha$ , power, and sample size”. The number of groups (corresponding here to naming conditions) was 4 and the number of measurements (corresponding here to the number of items in the naming task) was 128. This approximation takes into account the limitations of G\*Power, in which computing power for linear mixed effects models or repeated measures ANOVA with more than one within-subjects factor is not supported.
- An additional analysis calculated  $d'$  and C. Results were equivalent to the logistic regression and can be seen in the Appendix.

## Acknowledgement

We thank Jeroen van Paridon for his assistance in creating some of the stimuli used in this experiment, Caitlin Decuyper for her assistance in programming the experiments, and Annelies van Wijngaarden and the student assistants of the Psychology of Language department for their help with translations.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Funding

Open access funding provided by the Max Planck Society.

## ORCID

Eirini Zormpa  <http://orcid.org/0000-0002-6902-7768>

## References

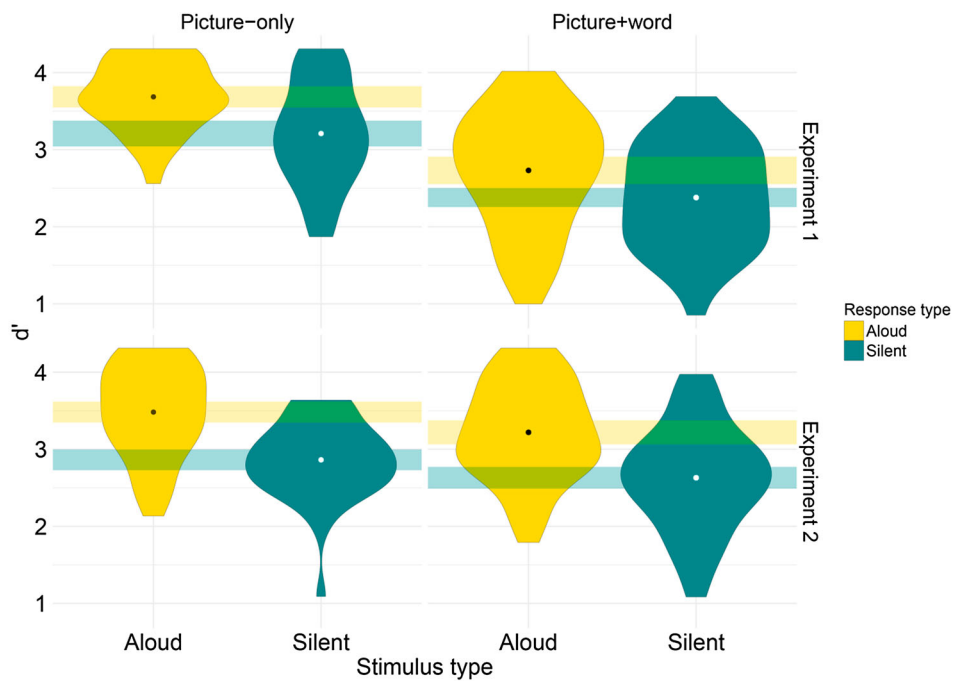
- Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01
- Begg, I., Snider, A., Foley, F., & Goddard, R. (1989). The generation effect is no artifact: Generating makes words distinctive. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(5), 977–989.
- Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, 35(2), 201–210.
- Bodner, G. E., Jamieson, R. K., Cormack, D. T., McDonald, D. L., & Bernstein, D. M. (2016). The production effect in recognition memory: Weakening strength can strengthen distinctiveness. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 70(2), 93–98. doi:10.1037/cep0000082
- Boersma, P., & Weenink, D. (2016). Praat, doing phonetics by computer (Version 6.0.22).
- Borges, M. A., Stepnowsky, M. A., & Holt, L. H. (1977). Recall and recognition of words and pictures by adults and children. *Bulletin of the Psychonomic Society*, 9(2), 113–114.
- Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The bank of standardized stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS One*, 5(5), e10773. doi:10.1371/journal.pone.0010773
- Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of standardized stimuli (BOSS) phase II: 930 new normative photos. *PLoS One*, 9(9), e106953. doi:10.1371/journal.pone.0106953
- Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, 26, 341–361.
- Curran, T., & Doyle, J. (2011). Picture superiority doubly dissociates the ERP correlates of recollection and familiarity. *Journal of Cognitive Neuroscience*, 23(5), 1247–1262. doi:10.1162/jocn.2010.21464
- DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, 3(2), 186–205. doi:10.1037/1082-989X.3.2.186
- Dell, G., & O’Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition*, 42, 287–314. doi:10.1016/0010-0277(92)90046-K
- Dodson, C. S., & Schacter, D. L. (2001). If I had said it I would have remembered it: Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin & Review*, 8(1), 155–161.
- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G\*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. doi:10.3758/Bf03193146



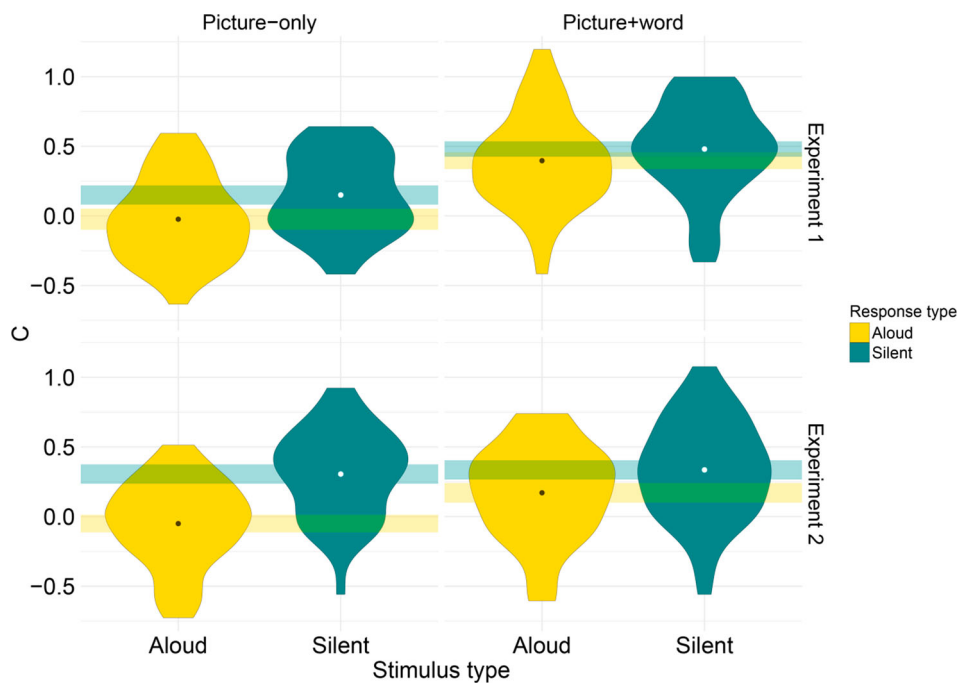
- Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subject production effect in recognition memory. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, doi:10.1037/cep0000089.supp
- Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory (Hove, England)*, 20(7), 655–666. doi:10.1080/09658211.2012.693510
- Forrin, N. D., Groot, B., & MacLeod, C. M. (2016). The d-prime directive: Assessing costs and benefits in recognition by dissociating mixed-list false alarm rates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(7), 1090–1111. doi:10.1037/xlm0000214
- Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory (Hove, England)*, 22(5), 470–480. doi:10.1080/09658211.2013.798417
- Forrin, N. D., & MacLeod, C. M. (2017). This time it's personal: The memory benefit of hearing oneself. *Memory (Hove, England)*, 26(4), 574–579. doi:10.1080/09658211.2017.1383434
- Forrin, N. D., Macleod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory & Cognition*, 40(7), 1046–1055. doi:10.3758/s13421-012-0210-8
- Fraundorf, S. H., Watson, D. G., & Benjamin, A. S. (2010). Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *Journal of Memory and Language*, 63(3), 367–386. doi:10.1016/j.jml.2010.06.004
- Gardiner, J. M., & Hampton, J. A. (1985). Semantic memory and the generation effect: Some tests of the lexical activation hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11(4), 732–741. doi:10.1037/0278-7393.11.4.732
- Gardiner, J. M., & Hampton, J. A. (1988). Item-specific processing and the generation effect: Support for a distinctiveness account. *The American Journal of Psychology*, 101(4), 495–504.
- Hoedemaker, R. S., Ernst, J., Meyer, A. S., & Belke, E. (2017). Language production in a shared task: Cumulative semantic interference from self- and other-produced context words. *Acta Psychologica*, 172, 55–63. doi:10.1016/j.actpsy.2016.11.007
- Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, 92(1–2), 101–144. doi:10.1016/j.cognition.2002.06.001
- Jacobs, C. L., Dell, G. S., Benjamin, A. S., & Bannard, C. (2016). Part and whole linguistic experience affect recognition memory for multiword sequences. *Journal of Memory and Language*, 87, 38–58. doi:10.1016/j.jml.2015.11.001
- Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, 42(3), 643–650. doi:10.3758/Brm.42.3.643
- Kinjo, H., & Snodgrass, J. G. (2000). Does the generation effect occur for pictures? *The American Journal of Psychology*, 113(1), 95–121.
- Kinoshita, S. (1989). Generation enhances semantic processing? The role of distinctiveness in the generation effect. *Memory & Cognition*, 17(5), 563–571. doi:10.3758/bf03197079
- Knutsen, D., & Le Bigot, L. (2014). Capturing egocentric biases in reference reuse during collaborative dialogue. *Psychonomic Bulletin & Review*, 21, 1590–1599. doi:10.3758/s13423-014-0620-7
- Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, 22, 1–75.
- MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review*, 18, 1197–1202. doi:10.3758/s13423-011-0168-8
- MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science*, 26(4), 390–395. doi:10.1177/0963721417691356
- MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(3), 671–685. doi:10.1037/a0018785
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). London: Erlbaum.
- Mama, Y., & Licht, M. (2016). Auditioning the distinctiveness account: Expanding the production effect to the auditory modality reveals the superiority of writing over vocalising. *Memory (Hove, England)*, 24(1), 98–113. doi:10.1080/09658211.2014.986135
- McKinley, G. L., Brown-Schmidt, S., & Benjamin, A. S. (2017). Memory for conversation and the development of common ground. *Memory & Cognition*, 45, 1281–1294. doi:10.3758/s13421-017-0730-3
- Nairne, J. S., Pusean, C., & Widner, R. L. (1985). Representations in the mental lexicon: Implications for theories of the generation effect. *Memory & Cognition*, 13(2), 183–191.
- Nairne, J. S., & Widner, R. L. (1987). Generation effects with nonwords: The role of test appropriateness. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(1), 164–171.
- Nelson, D. L., Reed, V. S., & McEvoy, C. L. (1977). Learning to order pictures and words: A model of sensory and semantic encoding. *Journal of Experimental Psychology: Human Learning and Memory*, 3(5), 485–497. doi:10.1037/0278-7393.3.5.485
- Ozubko, J. D., & Macleod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 36(6), 1543–1547. doi:10.1037/a0020604
- Paivio, A., Rogers, T. B., & Smythe, P. C. (1968). Why are pictures easier to recall than words? *Psychonomic Science*, 11(4), 137–138.
- Powell, M. J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives (*Cambridge NA Report NA2009/06*). University of Cambridge, Cambridge.
- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, 60(4), 437–447. doi:10.1016/j.jml.2009.01.004
- Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory (Hove, England)*, 21(8), 904–915. doi:10.1080/09658211.2013.766754
- Raven, J., Raven, J. C., & Court, J. H. (1998). *Raven manual section 4: Advanced progressive matrices*. Oxford, UK: Oxford Psychologists Press.
- R Development Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Richler, J. J., Palmeri, T. J., & Gauthier, I. (2013). How does using object names influence visual recognition memory? *Journal of Memory and Language*, 68(1), 10–25. doi:10.1016/j.jml.2012.09.001
- Rosson, M. B. (1983). From SOFA to LOUCH: Lexical contributions to pseudoword pronunciation. *Memory & Cognition*, 11(2), 152–160. doi:10.3758/BF03213470
- Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, 4(6), 592–604.
- Slamecka, N. J., & Katsaiti, L. T. (1987). The generation effect as an artifact of selective displacement rehearsal. *Journal of Memory and Language*, 26, 589–607.
- Smith, M. C., & Magee, L. E. (1980). Tracing the time course of picture-word processing. *Journal of Experimental Psychology: General*, 109(4), 373–392.
- Theios, J., & Muise, J. G. (1977). The word identification process in reading. *Cognitive Theory*, 2, 289–321.
- Weldon, M. S., & Roediger, H. L. 3rd. (1987). Altering retrieval demands reverses the picture superiority effect. *Memory & Cognition*, 15(4), 269–280.
- Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*, 41(2), 257–267. doi:10.3758/Brm.41.2.257
- Yoon, S. O., Benjamin, A. S., & Brown-Schmidt, S. (2016). The historical context in conversation: Lexical differentiation and memory for the discourse history. *Cognition*, 154, 102–117. doi:10.1016/j.cognition.2016.05.011



## Appendix



**Figure A1.** Sensitivity ( $d'$ ) results for each naming condition in Experiments 1 and 2. Columns represent stimulus type; rows represent experiment. Dots represent means by condition. Rectangles represent normalised within-subject 95% confidence intervals by condition.



**Figure A2.** Response bias ( $C$ ) results for each naming condition in Experiments 1 and 2. Columns represent stimulus type; rows represent experiment. Dots represent means by condition. Rectangles represent normalised within-subject 95% confidence intervals by condition.