Production of Picture Names Improves Picture Recognition

Kathleen L. Hourihan and Landon A. Churchill

Memorial University of Newfoundland

Author Note

Kathleen L. Hourihan, Department of Psychology, Memorial University of Newfoundland; Landon A. Churchill, Department of Psychology, Memorial University of Newfoundland.

This research was supported by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant to Kathleen Hourihan. We are grateful to Whitney Willcott-Benoit for her assistance with stimulus selection and to Victoria Kavanagh for her assistance with data collection and analysis.

Correspondence concerning this article should be addressed to Kathleen L. Hourihan, Department of Psychology, Memorial University of Newfoundland, St. John’s, NL, A1B 3X9, CANADA. Phone: (709) 864-8771. Email: khourihan@mun.ca

Abstract

Words read aloud are later recalled and recognized better than words read silently: the production effect. Previous research (Fawcett, Quinlan, & Taylor, 2012) has demonstrated a production effect in old/new recognition of line drawings. The current study examined whether production at encoding can improve memory for the visual details of a picture, or whether it is primarily memory for the picture's verbal label that benefits from production. Participants studied a list of photographs of nameable objects by naming half of the objects aloud and half silently. In Experiment 1, a control group completed a free recall test for the object names while the experimental group completed a 4-alternative forced-choice recognition test for the studied pictures and provided confidence judgements in their recognition decisions. Both groups showed a significant production effect. Experiment 2 obtained image typicality ratings and naming data for use in Experiment 3. In Experiment 3, studied items were tested after a one week delay in one of three different types of 2-alternative forced-choice recognition test: vs. a different picture exemplar of the same item; vs. a different picture; or as a verbal label vs. a different verbal label. Results showed a significant production effect in all testing conditions, with the magnitude of the effect similar across conditions. Production improves memory for both the visual details and verbal label of pictures.

Public Significance Statement: The production effect is the finding that when words are read aloud, they are later remembered better than words read silently. We found that when people say the names of pictures aloud, they remember the pictures better than if they say the names silently. The production effect appears to apply to memory for visual details.

Keywords: production effect, memory, recall, picture recognition

Production of Picture Names Improves Picture Recognition

The idea that words can affect the representations of objects is not novel. Verbal labels (e.g., names) activate perceptual features that are diagnostic of the *category* being labelled (Lupyan, 2008; 2012), but does saying the name of an object help us remember the *exact* visual details of that object? We were interested in examining the nature of the memory trace that is influenced by producing the name of an object at encoding: Does saying the name of an object improve memory for the visual details of that object, or does only memory for the object's verbal label benefit from production?

The production effect can be defined as the mnemonic advantage awarded to words that are produced (i.e., read aloud) relative to words that are not produced (i.e., read silently). Since being named by MacLeod, Gopie, Hourihan, Neary, and Ozubko (2010), research has shown that the memory benefits of production are not restricted to reading aloud: Words that are silently mouthed (MacLeod et al., 2010), typed, spelled, written (Forrin, MacLeod, & Ozubko, 2012) or sung (Quinlan & Taylor, 2013) are also better remembered, relative to words that are read silently. The memory advantage for produced items relative to silent items has been predominantly explained by a distinctiveness account (e.g., MacLeod et al., 2010; Ozubko & MacLeod, 2010; Ozubko, Major, & MacLeod, 2014). Although the effect is more often found within-subjects relative to between-subjects, multiple lines of research have found production effects using between-subject designs (e.g., Bodner & Taikh, 2012; Bodner, Taikh, & Fawcett, 2014; Fawcett, 2013; Fawcett & Ozubko, 2016), providing support for a strength account of the production effect. The goal of the current study was not to tease apart the relative contributions of strength and distinctiveness, but to examine the nature of the memory trace that is improved by production. To accomplish this, we examined how production influences memory for pictures.

It is well known that pictures are remembered better than their associated verbal labels: the picture superiority effect (Shepard, 1967). The picture superiority effect may be explained by dual coding theory (Paivio, 1971), or by the perceptual distinctiveness inherent in pictures (e.g., Ensor, Surprenant, & Neath, 2019). It has already been demonstrated that production can improve the recognition of pictures when the stimuli consist of line drawings. Fawcett, Quinlan, and Taylor (2012) were primarily interested in examining the possibility of an interaction between the production effect and the picture superiority effect, such that they expected the magnitude of the production effect (based on retrieving distinctive information) to be increased when memory is already based on distinctive information—for pictures relative to words. In a series of three experiments, participants were asked to study lists of line drawings and words (which consisted of the verbal labels of drawings) by producing half the items (i.e., mouth silently either the word or the picture's verbal label) and studying the other half silently. Regardless of whether study lists were blocked by stimulus type or mixed, or whether the recognition test was speeded or self-paced, production was found to improve memory for both pictures and words, and the predicted interaction was also observed. That is, the production benefit was relatively larger for pictures than for words. Thus, when pictures have an unambiguous verbal label that can be produced, memory for pictures is improved by production at encoding. The production benefit for naming pictures has also been replicated in young children (Icht & Mama, 2015), using both free recall and four-alternative-forced-choice recognition (using unmatched distractors).

Recently, research has focused on identifying boundary conditions of the production effect. In a series of experiments, Putnam, Ozubko, MacLeod, and Roediger (2014) found that the production effect was evident in associative memory, such that memory for word-word pairs was better if the pairs were spoken aloud rather than read silently. However, the effect was eliminated if participants made a semantic relatedness judgment on the pair following production. Hourihan and Smith (2016) subsequently examined whether production would aid in associative memory of face-name pairs. In their first experiment, participants were shown faces paired with a name, and were asked to either say the name aloud or read the name silently. At test, participants completed a cued-recall test in which they were asked to recall the name when prompted with the face. Experiment 2 followed the same procedure, with the exception that participants were asked to repeat the name aloud for the entire duration that the face was presented. No production effect was found in either experiment, leading the researchers to postulate that perhaps the associative production effect requires both members of a pair to be produced; that is, although participants could produce the names, they could not produce the faces, and thus a unitized word-image association was not created at encoding. Consistent with this theory, there was also no production effect found in cued recall when participants produced only one word from a word-word pair at encoding (Experiment 3).

The lack of a production effect in face-name associative memory has been explained by the inability to produce a face; that is, both items in a pair must be produced in order to benefit from production. However, when pictures have an unambiguous verbal label that can be produced at encoding, picture recognition does benefit from production at encoding (Fawcett et al., 2012). But does production benefit picture recognition by improving memory for the visual details inherent in the image, or is the boost in performance simply due to improved memory for that which was actually produced at encoding—the verbal label?

There is emerging evidence that producing a word does more than enhance memory for the produced verbal information. Fawcett, Blanchette, Rose, and Bodner (2018) examined whether production would influence the ability to discriminate between members of homophone pairs in a two-alternative forced-choice recognition test. For example, participants studied "ate", and at test were asked to discriminate "ate" from "eight". By definition, homophone pairs share the same phonological information, so if reading a word aloud at study only enhances memory for that particular phonology, then it should not help to discriminate between two words with identical phonology. However, Fawcett et al. found a production effect for words tested against matched homophone lures, the magnitude of which was similar to the production effect for words tested against unmatched lures, and in both between- and within-subject designs. Thus, production appears to enhance memory for more than the basic phonological information produced, allowing discrimination of targets from distractors that share phonological information.

Interestingly, these results conflict with those of Richler, Palmeri, and Gauthier (2013). Their experiments were designed to further test Lupyan's (2008) representational shift account of the influence of naming on object recognition. Specifically, Lupyan proposed that naming an object biases memory toward a prototype, reducing accurate memory for perceptual details of the specific exemplar (see also Lupyan, 2012), relative to a preference rating encoding task. However, Richler, Gauthier, and Palmeri (2011) demonstrated that the memory difference may be driven by an improvement caused by rating items, rather than a decrement due to naming items. In Richler et al.'s (2013) follow-up study, they explored the similarities between object naming (categorization) effects and production effects in memory for objects. Their experiments compared the combined effects of the task (naming vs. preference), with the uniqueness of items, the uniqueness of responses, and whether responses were verbal vs. keypress. They found production effects in old/new recognition, but only when unrelated lures were used at test. When lures were matched exemplars to study items, overall recognition performance was worse and there was no production effect.

This brings us to our current investigation. It may be that producing the name of an object will aid in correct recognition of that object; however, it may be only the verbal label of objects that benefits from production. Put simply, will saying “car” help to remember the *exact* car seen, or will it simply strengthen the memory for the “car” object category (cf. Lupyan, 2008)? In the current study, we examined whether producing the names of objects at encoding would result in better recognition of the exact picture of that object, relative to recognition for objects that were studied silently. In Experiments 1 and 3, participants studied pictures of nameable objects (e.g., brush, stove, apple) and were asked to say the names of some pictures aloud and others silently. We tested memory using forced choice recognition, asking participants to select the studied picture (e.g., an apple) when faced with similar exemplar pictures that share a common verbal label (e.g., pictures of different apples). (Experiment 2 obtained picture typicality ratings and naming data.) If production only benefits memory for that which is actually produced—the verbal label—then we would not expect to see any production effect when pictures are tested against pictures with the same verbal label. This result would fit with Richler et al.’s (2013) findings of no production benefit for objects when distractors are matched to study images. However, if producing an object’s name serves to enhance memory for perceptual details in addition to phonological details, then naming an object aloud at study may facilitate picture recognition, relative to naming an object silently.

Experiment 1

In this experiment, we included a group of participants who completed a free recall test for the picture names, to ensure that this study procedure would result in a production effect for the object names. A free recall test, rather than picture (or picture name) recognition was chosen to avoid ceiling performance, given that only 40 items were presented at study. Critically, the experimental group completed a four-alternative forced-choice (4-AFC) recognition test. Here, participants were asked to identify the picture they remembered from the study phase when provided with the target and three matched lures (similar pictures of different exemplars from the same object category).

In line with previous findings of the mixed-list production effect in recall (e.g., Icht & Mama, 2015; Lin & MacLeod, 2012; Mama & Icht, 2016), we expected that those who completed the free recall test would remember more names of objects studied aloud relative to objects studied silently. Importantly, if production improves recognition memory for the visual details of pictures, we expected that those who completed the 4-AFC recognition test would correctly recognize more pictures that they studied aloud relative to those studied silently. Conversely, if production at encoding improves memory only for that which is actually produced—the verbal label—then remembering saying “stove” aloud at study will do little to help differentiate among four similar images of stoves. In this case, we would expect no differences in recognition accuracy of pictures based on production, as found by Richler et al. (2013) using old/new recognition with matched lures.

Method

Participants. Sixty-four undergraduate students at Memorial University of Newfoundland participated for one course credit, or $10 CAD. Participants were pseudo-randomly assigned to group (recall or recognition) based on their arrival to the laboratory. Four participants were not included in data analyses due to failure to follow instructions (e.g., producing the incorrect items during study), leaving 30 participants in each group; there were no additional exclusion criteria. Demographic information was not collected.

Materials. The experimental procedure was presented using E-Prime v2.0 software (Psychology Software Tools Inc.) on a desktop computer running Windows 7. The stimuli consisted of 160 two-dimensional JPEG images that were selected from Brady, Alvarez, Konkle, and Oliva (2010), and Brodeur, Dionne-Dostie, Montreuil, and Lepage (2010). Each image file was presented in colour on a white background, and images within the same category were oriented in the same direction. All exemplars in a given category were the same colour (e.g., all apples were red; all peppers were green; all stoves were white, etc.). The on-screen picture dimensions at study and at test were fit to 35% x 40% of the screen. The 160 images were divided into 40 object categories, each consisting of four similar exemplar pictures (e.g., four stoves, four apples, four brushes, etc.); see Figure 1 for an example. For each participant, one of the four exemplar images was randomly selected to serve as the target and the remaining three images served as matched lures at test. At study, target images were presented in the center of the screen with either a red or blue border. The border colour served as production instructions and was randomly assigned to items across trials, resulting in 20 Aloud and 20 Silent items. During the 4-AFC recognition test, four images were presented in the four quadrants of the screen (one target and three lures), with the absence of colored borders. Each image was associated with a number from 1 to 4, such that participants could press the corresponding key on the keyboard to indicate the image they believed to be the target. The target image appeared in each location equally often, and the locations of the specific lure exemplars were randomly determined on each trial.

Procedure. Each participant completed a study phase, a distractor phase, and test phase (free recall or 4-AFC recognition). Participants were instructed that they would be studying a list of pictures for an upcoming memory test, with the goal to remember as many pictures as possible. Additionally, participants were informed that the colour of the border around each picture would indicate whether they should say the name of that picture aloud (blue) or silently (red).

Prior to the study phase, participants completed four practice trials to ensure that they understood the production instructions. Each study trial began with a white screen with a fixation point (+) in the center, presented for 500ms, followed by the study picture. Each picture was presented in the center of the screen for 2000ms and was followed by a blank screen for another 2000ms. A microphone was placed on the desk in front of the participant, and an experimenter remained in the room to monitor whether participants followed the encoding instructions correctly. Participants studied 40 pictures in total. Following the study phase, participants completed a distractor task for three minutes. This task required participants to solve a series of anagrams, by typing the appropriate response and pressing 'ENTER' on the keyboard.

Following the distractor phase, each participant completed either a free recall test or a 4-AFC recognition test. For the recall test, participants were asked to type in the names of as many pictures as they could remember, regardless of whether they read the name aloud or silently, by typing their response and pressing ‘ENTER’ on the keyboard. For the 4-AFC recognition test, each trial presented a studied target picture, along with three matched lures. Regardless of whether the image was studied aloud or silently, participants were asked to indicate which of the images they remembered from the study phase by pressing a number from 1 to 4 that corresponded to the picture on the screen. After each test trial, participants were also asked to indicate how confident they were that the picture they selected was the picture they studied, by pressing a number from 1 (unsure) to 4 (definitely sure) on the keyboard. All 40 studied pictures were tested, in random order. All responses were self-paced.

Results

Memory Performance. Mean correct performance in the recall and recognition groups is shown in the top portion of Table 1. Free recall performance was scored liberally (i.e., spelling and typographic errors were counted as correct). There was a significant production effect in the recall group, $t(29) = 2.72$, $p = .011$, $d = 0.496$. There was also a significant production effect in the recognition group, $t(29) = 5.23$, $p < .001$, $d = 0.954$. Recognition accuracy in both conditions was significantly above chance (i.e., 25%); both *p*s < .001. The confidence ratings provided by participants in the recognition group were significantly higher for responses to pictures that had been named aloud compared to those that had been named silently, $t(29) = 6.67$, $p < .001$, $d = 1.22$. The pattern remained the same when analyzing correct responses only, $t(29) = 4.00$, $p < .001$, $d = 0.73$ (see the bottom portion of Table 1).

Discussion

In this experiment, participants studied pictures by naming them aloud or silently at the time of encoding. Our control group showed that this study procedure resulted in a production effect in free recall. Importantly, the recognition group also showed a production effect when completing 4-AFC recognition. This result is inconsistent with the idea that production only enhances memory for that which is produced—the verbal information. If only verbal information were enhanced in memory, then there is no reason why remembering having said "apple" aloud during study would influence the ability to correctly select the specific studied exemplar when faced with three very similar alternative exemplars, all sharing the same verbal label. Instead, these results are consistent with the idea that the act of production enhances memory for the entire encoding episode. That is, the specific visual information present in the target picture appears to have been enhanced in memory by the act of naming it aloud, relative to naming it silently.

Experiment 2

An uninteresting explanation for the observed production effect in recognition would be if, by chance, the items assigned to be named aloud happened to be more typical than the items assigned to be named silently, and therefore were more likely to be selected at recognition regardless of whether the specific exemplar was actually remembered, consistent with Lupyan's (2008) representational shift hypothesis. Therefore, the primary goal of Experiment 2 was to obtain ratings of typicality for the items used in Experiment 1; performance was analysed on the basis of whether item typicality differentially influenced performance for aloud vs. silent items. In addition, the ratings were used to select pairs of items for Experiment 3 that were closely matched in terms of typicality, and to ensure that the verbal label of each item (to be presented in Experiment 3) was consistently provided by the majority of participants. Experiment 2 consisted of an online study in which participants provided a name and typicality ratings for 100 sets of four images, including the 40 categories used in Experiment 1 and 60 additional categories.

Method

Participants. The survey was initiated by 100 participants, who were undergraduate students at Memorial University of Newfoundland. Participants received one bonus course credit for completing the survey; none had participated in Experiment 1. Of the 100 participants, three did not complete the survey and two declined to have their data used for research (a required option for the credit participant pool). Of the remaining 95 participants, three were removed for failure to follow instructions (i.e., providing a ranking of 1 or 2 for all four exemplars the category, for multiple categories). The mean typicality ratings (and naming data) were computed using the responses from the remaining 92 participants.

Materials. The study was conducted online using the internal survey feature of Sona (Sona Systems Ltd.). Participants were shown 400 images in total, consisting of 100 sets of four different images depicting different but similar exemplars of an item (e.g., four images of plates). The images included all 160 of the items used in Experiment 1 (i.e., 40 sets of four images), additional items from the same sources, and additional items obtained from freedigitalphotos.net.

Procedure. Participants first saw an example page showing four images of different dogs (which were not included in the experimental set), and were instructed to first provide a name for the group of items (e.g., "dog"). They were asked to rank order the four items in terms how well they represented the name they had provided (i.e., typicality), with the best item being ranked "1" and the least representative item being ranked "4". Each set of four images was presented on a different page of the survey. Participant typed their category name in a text box, and selected their rating by clicking on a radio button. Only one number option could be selected for each picture, but the system did not prevent participants from responding with the same rank for more than one image (although they were instructed to use each number once, for one picture). Participants also had the option of declining to respond to any question. The 100 pages of the online survey (i.e., the order of the categories) were presented in a random order for each participant, and but the four exemplars images were always presented in the same order on the page (due to limitations in question ordering options within the survey system).

Results

Object naming consistency. In most cases, the names provided by participants matched the intended target name, although a minority of individual responses were more specific (e.g., “tote bag” instead of just “bag”; “storage locker” instead of just “locker”, “conch shell” instead of just “shell”, etc.) or less specific (e.g., “dessert” instead of “cake”; “musical instrument” instead of “accordion”; “fruit” instead of “plum”, etc.) than the target, or provided a common synonym (e.g., “sofa” instead of “couch”; “skillet” instead of “pan”, etc.). However, overall agreement on the names of most items was high. Names were considered to match the target name if they matched exactly or were more specific than the target, but still contained the target itself. For example, for the target “rose”, the responses “rose” and “red rose” would both be considered to match, but “flower” would not.

Averaging across all 100 items, the mean proportion of participants who provided the dominant name that was used as the verbal label in Experiment 3 was .922 (*SE* = .012; Min = .473; Max = 1.0; Mdn = .978; Mode = 1.0). For two of the 100 items, the verbal label used in Experiment 3 was altered based on participant responses: “beans” was changed to “coffee beans” (noting that no other type of bean was present in the pool), and “bottle” was changed to “water bottle” (again, there was no other type of bottle present in the pool). Even for the few items with lower rating agreement among names (e.g., “stove”/“oven”; “compass”/“protractor”; “cake”/“dessert”, “flowerpot”/“pot”) the label that was ultimately used in the experiment was sufficiently clear to discriminate the intended target from all other items in the pool, and aligned with the meaning provided by participants whose response did not match closely enough to be considered as the target. For example, although the item "flowerpot" had the lowest proportion of participants who provided the target name (.473), including participants who provided the name "pot" increased the proportion agreement on the name to .89. Similarly, if both "beans" and "coffee beans" were counted as consistent with the target "coffee beans", then agreement increased from .53 to .97. In both cases, the names used were more specific than the response provided by a significant number of participants, but unambiguously referred to the intended target pictures.

Typicality ratings and Experiment 1 performance. We first examined whether the mean typicality rating of the items randomly selected to serve as targets in Experiment 1 differed for Aloud vs. Silent items. The mean typicality (on a scale of 1 – 4, with lower numbers indicating higher typicality) of Aloud items was 2.47 (SE = 0.02), which did not differ from the mean typicality of Silent items (M = 2.53, SE = 0.02; $t(29)$ = 1.73, p = .095, d = 0.315). Thus, the production effect in recognition observed in Experiment 1 does not appear to be due random assignment of more typical exemplars to the Aloud condition than to the Silent condition.

We next used these mean typicality ratings to examine whether the errors made in recognition tended toward selecting an exemplar that was more typical than the studied exemplar, and whether this tendency differed for aloud vs. silent trials. (Note that two participants did not make any errors with Aloud items and therefore were not included in the analysis.) That is, when an error was made, were typical exemplars more likely to be selected when the item had been named aloud at study than when it had been named silently? The mean typicality rating of incorrect choices on Aloud trials was 2.32 (SE = 0.05), which did not differ from the mean typicality of incorrect choices on Silent trials (M = 2.32, SE = 0.04; $t(27)$ = 0.10, *p* = .924, *d* = 0.018). When comparing the typicality of the specific exemplar chosen on an error trial to the typicality of the actual target (where positive numbers indicate a selection more typical than the target and negative numbers indicate a selection less typical than the target), participants were likely to select an exemplar that was more typical (rather than less) than the target, but were equivalently so on Aloud (*M* = 0.23, *SE* = 0.09) and Silent trials (*M* = 0.25, *SE* = 0.07; *t*(27) = 0.22, *p* = .829, *d* = 0.04). Thus, although exemplar typicality did influence participants' choices when they did not select the correct target, there was no systematic difference based on production.

Discussion

Experiment 2 obtained exemplar typicality ratings and naming consistency data for the images used in Experiment 1 as well as a large number of additional image categories. Primarily, we showed that the production effect observed in Experiment 1 was not an artefact of the confounding of item typicality and assignment to production condition. That is, the specific exemplars named aloud were not more typical than those named silently, and although errors indicated that participants were likely to choose a more typical exemplar than the target (cf. Lupyan, 2008), this occurred similarly in the two conditions. The typicality ratings obtained in Experiment 2 were used to select pairs of items that were similar in typicality for use in Experiment 3.

Experiment 3

In Experiment 1, we found that producing a picture's name at the time of encoding improved the ability to discriminate a specific exemplar image from similar distractor images. In Experiment 3, our goal was to compare the magnitude of this production effect to other testing conditions. Considering the test conditions in Experiment 1, the target item could only be correctly recognized on the basis of remembering specific visual features of the specific exemplar—the distractor items all shared the same verbal label, so remembering that "apple" was said aloud at study should do little to help discriminate the target from distractors. In Fawcett et al.'s (2012) study examining the production effect for pictures, they argued that the production effect and the picture superiority effect may both operate on the basis of distinctive cues, such that the presence of multiple sources of distinctive information (i.e., a picture's visual details and the produced verbal information) interact to enhance memory. They showed a larger magnitude production effect for pictures than for words across three experiments. In all of their experiments, the memory test was old/new recognition, in which recognition could be enhanced (relative to silent words) on the basis of remembering either the distinctive visual image or the distinctive production information; remembering both led to an especially distinctive memory. We have shown in Experiment 1 that memory for pictures is enhanced by production even when the produced information is not helpful for target discrimination—would the magnitude of the production effect for images be further enhanced when both the visual and verbal information can differentiate target from distractor?

In Experiment 3, we further examined the relative benefit of production on picture memory, using a 2-AFC recognition test[1]. One third of the studied items were tested as in Experiment 1, with the distractor item being another exemplar of the same item studied (exemplar test). One third of the studied items were tested with the distractor item being a picture of a different item not encountered at study (picture test). The remaining items were tested as words—the verbal label of the studied picture—with the distractor item being another word, not

[1] We used a 2-AFC test rather than 4-AFC test (as in Experiment 1) primarily due to challenges in finding sufficient numbers of items (with each "item" consisting of a set of four visually similar picture stimuli) and maintaining the ability to randomly assign all items to conditions (see Materials section below). That is, our current pool used 100 items, but for 4-AFC, we would have required an additional 80 items in the pool.

encountered at study (word test). A production effect was expected for all tests. The magnitude of these production effects, however, was expected to differ on the basis of the test. Specifically, we expect the largest magnitude production effect in the picture test, in which the target can be discriminated from the distractor on the basis of both visual and verbal information (similar to the larger magnitude production effect observed by Fawcett et al., 2012). In the exemplar test, the verbal information would apply to both target and distractor, and thus the target could only be correctly identified on the basis of remembering its specific visual details. In the word test, the visual information is not reinstated, but remembering the produced verbal information can clearly discriminate the target from the (equally imageable) distractor.

Method

Participants. Twenty-seven undergraduate students at Memorial University of Newfoundland participated for two course credits (one per session). None had participated in the first two experiments.

Materials. The experimental procedure was presented using E-Prime v3.0 software (Psychology Software Tools Inc.) on a desktop computer running Windows 7. Based on the mean typicality ratings obtained in Experiment 2, the two exemplars of a given item that were the closest in mean rating were selected for the item pool. If there was a tie in terms of the smallest difference in ratings, then the pair of items with the lowest variability (across raters) in rating were selected. The stimuli therefore consisted of 200 two-dimensional JPEG images, including two exemplars of each object, for a total of 100 items.

Images were displayed as in Experiment 1. For each participant, a random 60 items were selected to serve as study items; the specific exemplar that was shown at study was randomly selected. Of the remaining 40 items in the pool, 20 were selected to serve as distractors for the picture test trials and 20 served as distractors for the word test trials. In the exemplar test trials, studied targets were paired with the other image exemplar of that item from the pool that had not been shown at study. In the picture trials, studied targets were paired with another image of a different exemplar from the pool that had not been shown at study, and the specific exemplar shown was randomly selected. For the word trials, targets were tested as the verbal label of the studied image (based on the names obtained in Experiment 2), paired with the verbal label of a different exemplar from the pool that had not been shown at study.

Procedure. Each participant completed a study phase and returned to the lab one week later to complete the recognition test[2]. The study phase was nearly identical to Experiment 1, except that there were 60 trials (30 Aloud and 30 Silent) instead of 40.

The recognition test consisted of 60 test trials: 20 exemplar test trials (10 Aloud and 10 Silent), 20 picture test trials (10 Aloud and 10 Silent), and 20 word test trials (10 Aloud and 10 Silent), in random order. Regardless of whether the image was studied aloud or silently, participants were asked to indicate which of the two items they remembered from the study phase by pressing the “z” key to indicate the item on the left of the screen and the “m” key to indicate the item on the right of the screen. Targets appeared on the left and the right equally often. As in Experiment 1, participants were also asked to indicate how confident they were that the picture they selected was the picture they studied, by pressing a number from 1 (unsure) to 4 (definitely sure) on the keyboard. All responses were self-paced.

Results

Recognition Accuracy. Mean correct recognition performance is shown in Table 2. Correct recognition scores were analyzed with a 2 (production: aloud vs. silent) x 3 (test type:

[2] Pilot testing showed performance at ceiling when tested in the same session as the study phase.

exemplar vs. picture vs. word) repeated measures analysis of variance (ANOVA). The main effect of production was significant, $F(1,26) = 57.73$, $MSE = 0.012$, $p < .001$, $\eta_p^2 = .689$, with greater recognition accuracy for pictures named aloud ($M = .79$, $SE = .014$) compared to pictures named silently ($M = .66$, $SE = .02$). The main effect of test type was also significant, $F(2, 52) = 32.28$, $MSE = 0.014$, $p < .001$, $\eta_p^2 = .554$. Post-hoc t-tests (with Tukey correction) showed that performance on the picture test ($M = .83$, $SE = .018$) was better than on the exemplar test ($M = .69$, $SE = .021$, $t(52) = 5.85$, $p < .001$) and the word test ($M = .65$, $SE = .020$, $t(52) = 7.69$, $p < .001$), but the exemplar test and word test did not differ from one another, $t(52) = 1.84$, $p = .166$. Contrary to predictions, the production x test type interaction was not significant, $F(2,52) = 0.22$, $MSE = 0.019$, $p = .802$. Performance in all conditions was significantly above chance; all ps $\leq .017$.

Given the surprising null interaction between test type and production, Bayesian analyses were conducted on the interaction, using the approximation provided by Masson (2011). The resulting posterior probability in favour of the null hypothesis was $p_{BIC}(H_0|D) = .96$, therefore showing strong evidence that the magnitude of the production effect did not differ across the three test types.

Recognition Confidence. Mean recognition confidence ratings are shown in Table 2. Mean confidence ratings were analyzed in a 2 (production: aloud vs. silent) x 3 (test type: exemplar vs. picture vs. word) repeated measures ANOVA. The main effect of production was significant, $F(1,26) = 32.76$, $MSE = 0.152$, $p < .001$, $\eta_p^2 = .558$, with higher confidence ratings for pictures named aloud ($M = 2.73$, $SE = 0.09$) compared to pictures named silently ($M = 2.38$, $SE = 0.10$). The main effect of test type was also significant, $F(2, 52) = 33.89$, $MSE = 0.139$, $p < .001$, $\eta_p^2 = .566$. Following the same pattern as recognition accuracy, post-hoc t-tests (with

Tukey correction) showed that confidence ratings on the picture test (*M* = 2.90, *SE* = 0.09) were higher than on the exemplar test (*M* = 2.38, *SE* = 0.10, $t(52) = 7.27$, $p < .001$) and the word test (*M* = 2.40, *SE* = 0.12, $t(52) = 6.98$, $p < .001$), but the exemplar test and word test did not differ from one another, $t(52) = 0.28$, $p = .957$. The production x test type interaction was not significant, $F(2,52) = 1.21$, *MSE* = 0.078, $p = .307$. The same pattern of results was observed when analyzing confidence of correct responses only (see Table 2).

Discussion

The goal of Experiment 3 was to examine whether the magnitude of the production effect for pictures would vary depending on the way in which the target was tested. Replicating the results of Experiment 1, a production effect was found when targets were tested with another exemplar of the same item as the distractor (exemplar test). A production effect was also found when targets were tested with a different picture as the distractor (picture test) and when the target was tested as its verbal label against the label of a different item as the distractor (word test). Contrary to predictions, the magnitude of the production effect did not differ across the three tests. We did, however, observe better performance on the picture test than on the other two test types (which did not differ from one another).

As described above, the three test types differ on the nature of the information that must be remembered in order to select the target. For the exemplar test, the visual details of the specific exemplar studied must be recognized in order to differentiate the target from the distractor. Similarly, for the word test, the verbal label must be recognized in order to differentiate the target from the distractor. For the picture test, however, remembering either the visual details or the verbal label would be sufficient to discriminate target from distractor; it had been predicted that, as in Fawcett et al. (2012), remembering both pieces of information would result in an additional memory benefit, and enhance the production benefit. Interestingly, we did not see any evidence that the production effect was enhanced in the picture test, although overall performance was best in this condition. That is, both aloud and silent pictures were better recognized when tested against a distractor that differed in both visual details and verbal label, compared to when tested against distractors only differing in one piece of information. This differs from the result found by Fawcett et al., who observed a larger magnitude production effect for pictures compared to words.

One key difference between Fawcett et al.'s (2012) study and the current is the nature of the recognition test; whereas the current experiment used 2-AFC, Fawcett et al. used old/new recognition tests. In old/new recognition (but not in 2-AFC), there is potential for response bias to influence recognition performance; indeed, Fawcett et al., reported a significantly more liberal response bias for aloud items than for silent items, along with a smaller effect of response bias for pictures (more liberal) compared to words. Rather than comparing memory for a single item to a response criterion, forced-choice recognition relies on a relative judgment between the two items. However, although the two types of recognition tests have been shown to have similar levels of accuracy (e.g., Kroll, Yonelinas, Dobbins, & Frederick, 2002), forced choice recognition is considered to be an easier test than old/new recognition (e.g., Jang, Wixted, & Huber, 2009; MacMillan & Creelman, 2005). Thus, the easier decision process involved with forced choice recognition (compared to old/new recognition) may have negated any additional benefit of production when recognizing pictures compared to the verbal label of those pictures.

General Discussion

The current study examined the nature of the information that is improved by production at encoding: Can production improve memory for the visual details of pictures, or does it only improve memory for a picture's verbal label? Participants were shown pictures of nameable objects and were asked to either produce the names of the objects, or to name the objects silently. In two experiments, production at study improved forced-choice recognition of the target object, even when tested against distractors that shared the same verbal label—different exemplars of the same object. Thus, the current study provides evidence that the act of production improves memory for more than that which is produced—the verbal label—and indicates that perceptual information is also enhanced.

Exactly how does production improve memory for the specific visual details of images? It may be the case that when a picture name is produced aloud at study it enhances the association between that verbal label and the specific visual features in the target. At recognition, the visual details of the target are reinstated, which may lead to recollection of the original study episode. Previous research with words has shown that production (in mixed-list designs) enhances recollection in addition to familiarity (Fawcett & Ozubko, 2016; Ozubko, Gopie, & MacLeod, 2012); participants are more likely to remember additional information about the study episode (e.g., whether the item was studied aloud or silently; Ozubko et al., 2012; 2014) when the target was produced. Even in the absence of recollection, the studied target may feel more familiar. Or, even if episodic memory of studying the target were absent, then at recognition the associated verbal label would be readily retrievable from semantic memory (i.e., the object could be named). If production enhances the association between the verbal label and the target's specific visual details, then memory for the visual details of the target may be cued simply by naming the object. Thus, if production enhances the link between the label and specific visual details, then recognizing either one would make it easier to retrieve the associated information.

But how do the current findings fit with the conclusions from Hourihan and Smith's (2016) study demonstrating that face-name pairings do not benefit from production, most likely because faces cannot be produced? In the current study, the targets were all common objects that were readily and consistently named by participants (see Experiment 2). This means that the targets all had pre-existing entries in semantic memory, so the names of the targets only had to be retrieved, not learned, a process which is thought to operate relatively automatically (e.g., Richler et al., 2011). With face-name pairs, although the names were sufficiently common that participants would have been familiar with them (i.e., the names would exist in semantic memory), the faces were all novel faces from a database. Thus, the face-name pairing had to be learned, and only one part (the name) of the new associative pair could be produced. With the common objects in the current study, the pairing between visual and verbal information was not novel.

Speculatively, we can therefore make two predictions about related circumstances. First, if a face-name pairing were not novel, such that the person's name could be readily retrieved from semantic memory upon seeing the face, then production might benefit memory for face-name pairs. For example, if participants studied images of well-known relatives or friends by naming them aloud or silently at study (in the absence of a specific name label to read), then an old/new recognition test (or even a forced-choice test between two pictures of the same person) may show a production effect. Second, if objects were completely novel, such that participants had to learn the association between the name and the visual details (and could not simply retrieve the name from semantic memory), then production might not benefit memory for such pictures. That is, although single items do not need to have pre-existing semantics in order to benefit from production (e.g., non-words show production benefits, MacLeod et al., 2010; novel object learning in children benefits from production, Icht & Mama, 2015), it may be the case that for production to improve associative memory, the association itself must be the only new information learned in the episode, and not the identity of either member of a new associative pair.

The current study demonstrated that production at encoding improves memory for more than that which is produced—verbal information—extending the memory benefit to associated but unproduced information (i.e., visual information). Our results do not differentiate between the strength (e.g., Bodner et al., 2013) and distinctiveness (e.g., MacLeod et al., 2010) accounts of the production effect, nor were they intended to do so. Instead, our findings speak to the nature of the mnemonic information that is either strengthened or made distinct by production at encoding. Production not only improves memory when testing is reliant on the ability to retrieve verbal information, but also when the target's integrated visual details must be discriminated from similar distractors.

References

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology*, *38*, 1711-1719. doi: 10.1037/a0028466

Bodner, G. E., Taikh, A., & Fawcett, J. M. (2014). Assessing the costs and benefits of production in recognition. *Psychonomic Bulletin and Review*, *21*, 149-154. doi: 10.3758/s13423-013-0485-1

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2012). Real-world objects are not represented as bound units: Independent forgetting of difference object details from visual memory. *Journal of Experimental Psychology: General*, *142*, 791-808. doi: 10.1037/a0029649

Brodeur MB, 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

E-Prime (version 2.0; 3.0) [Computer Software]. Pittsburgh, PA: Psychology Software Tools, Inc.

Ensor, T.M., Surprenant, A.M., & Neath, I. (2019). Increasing word distinctiveness eliminates the picture superiority effect in recognition: Evidence for the physical-distinctiveness account. *Memory & Cognition*, *47*, 182-193. doi: 10.3758/s13421-018-0858-9

Fawcett, J. M. (2013). The production effect benefits performance in between-subjects designs: A meta-analysis. *Acta Psychologica*, *142*, 1-5. doi: 10.1016/j.actpsy.2012.10.001

Fawcett, J.M., Blanchette, T., Rose, J.P., & Bodner, G. E. (2018, July). *Did you study "ate" or "eight"? Homophone lures do not reduce the production effect in a 2AFC recognition* *task*. Poster presented at the annual meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science, St. John's, NL.

Fawcett, J. M., & Ozubko, J. D. (2016). Familiarity, but not recollection, supports the between-subjects production effect in recognition memory. *Canadian Journal of Experimental Psychology, 70,* 99-115. doi: 10.1037/cep0000089

Fawcett, J. M., Quinlan, C. K., & Taylor, T. L. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory, 20,* 655-666. doi: 10.1080/09658211.2012.693510. doi: 10.1080/09658211.2012.693510

Forrin, N.D., MacLeod, C.M., & Ozubko, J.D. (2012). Widening the boundaries of the production effect. *Memory & Cognition, 40,* 1046-1055. doi: 10.3758/s13421-012-0210-8

Hourihan, K. L., & Smith, A. R. S. (2016). Production does not improve memory for face-name associations. *Canadian Journal of Experimental Psychology, 70,* 147-153. doi: 10.1037/cep0000091

Icht, M., & Mama, Y. (2015). The production effect in memory: a prominent mnemonic in children. *Journal of Child Language, 42*, 1102-1124. doi: 10.1017/S0305000914000713

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *Journal of Experimental Psychology: General, 138*, 291-306. doi: 10.1037/a0015525

Kroll, N.E.A., Yonelinas, A.P., Dobbins, I.G., & Frederick, C.M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *Journal of Experimental Psychology: General, 131*, 241-254. doi: 10.1037/0096-3445.131.2.241

Lin, O.Y.H., & MacLeod, C.M. (2012). Aging and the production effect: A test of the distinctiveness account. *Canadian Journal of Experimental Psychology*, *66*, 212-216. doi: 10.1037/a0028309

Lupyan, G. (2008). From chair to “chair”: A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*, 248-369. doi: 10.1037/0096-3445.137.2.348

Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology, 3*. doi: 10.3389/fpsyg.2012.00054

Mama, Y., & Icht, M. (2016). Influence of retrieval mode on effects of production: Evidence for costs in free recall. *Canadian Journal of Experimental Psychology, 70,* 177-185. doi: 10.1037cep0000090

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*, 671-685. doi: 10.1037/a0018485

Masson, M.J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavioral Research Methods*, *43*, 679-690

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,* 1543-1547. doi: 10.1037/a0020604

Ozubko, J.D., Gopie, N., & MacLeod, C.M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition, 40,* 326-338. doi: 10.3758/s13421-011-0165-1

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

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart & Winston. Reprinted 1979, Hillsdale, NJ: Lawrence Erlbaum Associates.

Paladin13 (2015). Cactus stock photo. [Online image]. Retrieved from http://www.freedigitalphotos.net/images/cactus-photo-p376926.

Putnam, A. L., Ozubko, J. D., MacLeod, C. M., & Roediger, H. L. (2014). The production effect in paired-associate learning: Benefits for item and associative information. *Memory and Cognition, 42,* 409-420. doi: 10.3758/s13421-013-0374-x

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory, 21*, 904-915. doi: 10.1080/09658211.2013.766754

Richler, J.J., Gauthier, I., & Palmeri, T.J. (2011). Automaticity of basic-level categorization accounts for labeling effects in visual recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 1579-1587. doi: 10.1037/a0024347

Richler, J. J., Palmeri, T. J., & Gauthier, I. (2013). How does using object names influence visual recognition memory? *Journal of Memory and Language, 68*, 10-25. doi: 10.1016/j.jml.2012.09.001

Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning & Verbal Behavior, 6*, 156-163

Sona Systems [Computer Software]. Tallinn, Estonia: Sona Systems Ltd.

Stockdevil (2016). Cactus on isolated background (Cereus Hexagonus Mill). [Online image]. Retrieved from http://www.freedigitalphotos.net/images/cactus-on-isolated-background--cereus-hexagonus-mill--photo-p414656.

Stockdevil (2016). Cactus on isolated background (Cereus Hexagonus Mill). [Online image]. Retrieved from http://www.freedigitalphotos.net/images/cactus-on-isolated-background--cereus-hexagonus-mill--photo-p4146567.

Stockdevil (2016). Cactus (opuntia) on isolated background (Cereus hexagonus mil). [Online image]. Retrieved from http://www.freedigitalphotos.net/images/cactus--opuntia--on-isolated-background--cereus-hexagonus-mil-photo-p414655.

Table 1

Mean Performance Across Free Recall and 4-AFC Recognition Groups in Experiment 1

	Aloud	Silent
Proportion Correct		
Free Recall	.30 (.02)	.22 (.02)
4-AFC Recognition	.78 (.02)	.64 (.03)
Recognition Confidence		
Overall	3.09 (.08)	2.60 (.08)
Correct Responses Only	3.30 (.07)	3.01 (.07)

Note: Standard errors are in parentheses besides their respective means.

Table 2

Mean Performance on 2-AFC Recognition Tests in Experiment 3

	Aloud	Silent
Proportion Correct		
Exemplar Test	.77 (.02)	.62 (.03)
Picture Test	.89 (.02)	.77 (.02)
Word Test	.72 (.02)	.59 (.03)
Recognition Confidence		
Overall		
Exemplar Test	2.55 (0.10)	2.20 (0.11)
Picture Test	3.11 (0.10)	2.68 (0.10)
Word Test	2.53 (0.13)	2.26 (0.12)
Correct Responses Only		
Exemplar Test	2.67 (0.10)	2.46 (0.13)
Picture Test	3.19 (0.10)	2.83 (0.10)
Word Test	2.63 (0.13)	2.36 (0.13)

Note: Standard errors are in parentheses besides their respective means.

Figure 1. Example images used in Experiments 2 and 3 (similar to those used in Experiment 1). Images courtesy of users paladin13 and stockdevil at FreeDigitalPhotos.net.