

RUNNING HEAD: Naming and Recognition Memory

How does using object names influence visual recognition memory?

Jennifer J. Richler*, Thomas J. Palmeri & Isabel Gauthier

Vanderbilt University

***Corresponding Author:**

Jennifer J. Richler

Phone: 615 322 0229

Fax: 615 343 8449

Email: jennifer.j.richler@vanderbilt.edu

REGULAR MAIL (via U.S. Postal Service)

Vanderbilt University

PMB 407817

2301 Vanderbilt Place

Nashville, TN 37240-7817, USA

COURIER MAIL (via Fed Exp, UPS)

Department of Psychology

301 Wilson Hall

Vanderbilt University

Nashville, TN 37240, USA

ABSTRACT

Two recent lines of research suggest that explicitly naming objects at study influences subsequent memory for those objects at test. Lupyan (2008) suggested that naming *impairs* memory by a representational shift of stored representations of named objects toward the prototype (labeling effect). MacLeod and colleagues (MacLeod, Ozubko, Forrin & Hourihan, submitted; MacLeod, Gopie, Hourihan, Neary, & Ozubko, 2010) suggested that naming *enhances* memory by influencing the distinctiveness of named objects (production effect). However, these studies cannot be directly compared because they differ in several procedural details such as the format of the naming task, composition of study objects from different categories, control task, and type of lures used at test. Here we systematically manipulate those factors to better understand how using object names influences visual recognition memory. When objects belonged to unique categories, vocal naming (as used in the production effect) produced comparable memory as a non-naming task (preference rating) and both produced significantly better memory than key-press naming (as used in the labeling effect). Naming objects at study only impaired memory relative to preference rating when objects belonged to one of two categories, a condition in which names have little or no distinctiveness. Theoretically, our results pose challenges to the representational shift account that proposes special mechanisms tied to the use of object names.

KEYWORDS: naming; visual recognition memory

Language is critical for human communication and central to social interactions. To facilitate communication, object names are automatically accessed in preparation for speech (e.g., Meyer & Damian, 2007; Morsella & Miozzo, 2002; Navarrete & Costa, 2005). The seemingly automatic activation of object names has consequences for perceptual and cognitive processes, even outside of communicative goals. For example, gaze duration for pictures is correlated with name length and spoken name duration (Zelinsky & Murphy, 2000), objects that belong to the same linguistic category are perceived as being more similar than objects belonging to different linguistic categories (Goldstone, 1994; Roberson & Davidoff, 2000), and picture recall is influenced by the length and phonological similarity of picture names (Coltheart, 1999; Schiano & Watkins, 1981).

Having names for objects influences perception and memory. But is there an effect of *overtly* using object names? Many theories of object categorization (e.g., Ashby et al., 1998; Kruschke, 1992; Lamberts, 2000; Nosofsky & Palmeri, 1997) and object recognition (e.g., Joyce & Cottrell, 2004; Riesenhuber & Poggio, 1999; Serre et al., 2007) assume a bottom-up process where object representations are compared to category representations in order to name an object (Palmeri & Tarr, 2008). If overtly using object names systematically affects how objects are actually represented, this could have theoretical consequences for these theories.

Two recent lines of research have reported that intentionally and overtly naming objects during study can systematically influence subsequent visual memory for those objects at test. One reported impaired memory while the other reported enhanced memory. However, these lines of research are difficult to compare and integrate because their paradigms differ along many experimental factors. Our empirical goal is to systematically

investigate differences between these paradigms, allowing us to better understand how using object names influences recognition memory, which can have important theoretical implications. We describe these two lines of research in turn below.

Lupyan (2008) reported impaired recognition memory after naming objects at study. Specifically, memory was worse for objects labeled as chairs or lamps compared to objects rated for preference – a finding dubbed the *labeling effect* (Lupyan, 2008; see also Richler et al., 2011). Lupyan (2008) proposed a representational shift hypothesis to explain this impaired memory (see also Lupyan, 2012): Overtly naming objects exaggerates effects of object categorization, activating features associated with prototypical examples of the object's category. In a top-down manner, these features become coactive with the visual features of the named object and systematically alter the object representation stored in visual long-term memory. Overtly labeling a picture of a chair as a "chair" shifts its visual memory representation toward the chair prototype. The subsequent mismatch between a previously studied chair presented again at test and its "shifted" visual memory representation leads to a false sense that the previously studied object is a new chair, impairing memory performance.

Whereas representational shift predicts memory impairment from naming, in an independent line of research, MacLeod and colleagues reported enhanced memory from naming. Recognition memory was better for words (MacLeod et al., 2010) and line drawings of objects (MacLeod et al., submitted) named out loud than those named silently. This was dubbed the *production effect* (MacLeod et al., 2010) and was explained by MacLeod et al. in terms of distinctiveness: Vocal production of a name leads to more unique processing of the study item during encoding that can be "replayed" at test. A match

between processing at study and reprocessing at test facilitates recognizing that item as “old” (Conway & Gathercole, 1987; MacLeod, 2010; MacLeod et al., 2010; Ozubko & MacLeod, 2010). Alternatively, Dodson and Schacter (2001) proposed a metacognitive explanation for production effects. They suggested that saying words out loud at study reduces rates of false recognition because participants *expect* to remember having said a word out loud and the absence of this expected information is used as a cue that a test item is new. So unlike the labeling effect, the production effect has been explained without appealing to anything “special” about naming *per se*. Vocalizing the name of an object is just one of many things a person could do to make a memory representation more distinctive.

Critically, details of the experimental procedures used to test for the labeling effect (Blanco & Gureckis, 2011; Lupyan, 2008; Richler et al., 2011) and the production effect (MacLeod, 2010; MacLeod et al., 2010; MacLeod et al., submitted; Ozubko & MacLeod, 2010) differ in several important respects. Both use a relatively standard study-test procedure for recognition memory, but, as shown in Figure 1, the two paradigms differ on the exact nature of the naming task, the control task, and the composition of study objects from different categories (see also Figure 2). In the present article, we explore the missing cells in Figure 1, along with some not illustrated. Our goal was not merely to fill in a table of missing experimental conditions, but to explore key factors that might help elucidate the conditions under which a labeling effect (impaired memory from naming) or a production effect (enhanced memory from naming) can be obtained, with an eye toward a theoretical understanding of how using object names affects object representations and object memory.

First, the overt naming task differs between paradigms. In studies of the production effect, participants say the name of an object out loud (henceforth referred to as *vocal naming*). In studies of the labeling effect, participants are given a two-alternative forced choice key-press for the name of an object (henceforth referred to as *key-press naming*). In the following experiments we directly compare vocal naming and key-press naming. Theoretical accounts of the production effect predict that vocal naming leads to better memory than key-press naming, either because a vocal response leads to more distinctive memory than a key-press response (MacLeod et al., 2010) or because participants expect to remember a vocal response better than a key-press response (Dodson & Schacter, 2001). In contrast, representational shift does not make specific predictions, but is compatible with several different scenarios: On the one hand, the representational shift account could explain equivalent memory for vocal naming and key-press naming because representational shift occurs whenever category labels are explicitly activated and used. On the other hand, the representational shift account could explain worse memory for vocal naming leads compared to key-press naming because vocal naming is more overt, exaggerating the effects of categorization, leading to an even greater representational shift. Only by including both vocal naming and key-press naming in the same experiment can we discriminate between these possibilities.

Second, naming has been compared to different control tasks in studies of the production effect and labeling effect. For the production effect, overt naming was compared to silent naming. For the labeling effect, overt naming was compared to preference rating (“like” or “don’t like”). It is quite possible that overt naming improves memory relative to silent naming (the production effect), but both kinds of naming impair memory relative to

preference rating (the labeling effect). In other words, both effects of naming may be at work, but an experiment containing both kinds of control tasks is needed to observe them simultaneously. Theoretically, this would mean that both distinctiveness from overt compared to silent naming and representational shift from any kind of naming operate simultaneously.

Third, there are important differences in the composition of study objects from different categories. The labeling effect has been investigated with all objects from only two categories (e.g., chairs versus lamps), whereas the production effect has only been investigated with all objects come from a unique basic-level category. This experimental difference could be one source of the apparent discrepancy between the labeling effect and the production effect. For example, Lupyan (2008) hypothesized that naming might actually improve recognition memory for objects coming from different categories because the representational shift caused by naming would make the representation of each study item less noisy and more categorical. By this, a labeling effect would be predicted with only two categories, but no labeling effect, or even a “reverse” labeling effect (enhanced memory from naming), would be predicted with multiple unique categories. MacLeod and colleagues (MacLeod et al., 2010; Ozubko & MacLeod, 2010) suggested that unique names could be required to generate the relative distinctiveness needed to observe any memory benefits from naming objects vocally. By this account, a production effect would be predicted with multiple categories, but no production effect would be predicted with only two categories.

Finally, the nature of the lures used during in the old/new recognition memory test differs. For the production effect, new lures at test have been objects from other unstudied

basic-level categories that do not share the same name as any of the studied items (*unrelated lures*). For the labeling effect, new lures at test have been specifically matched to particular study items, with these *matched lures* differing from old objects in small but perceptible ways, such as the presence or absence of a feature or the height-to-width ratio. One advantage of matched lures is that they require participants to discriminate perceptually whether they have seen a particular object or not, without being able to use semantic or lexical information. In fact, if representational shift highlights the between-category difference between old and new items, then representational shift should predict improved memory by naming when lures come from unstudied categories (in contrast to the observed impairment when matched lures are used). The distinctiveness account does not predict any differences based on lure type.

In summary, we aim to create an empirical bridge between studies of the labeling effect and production effect. Ultimately, this is important for evaluating different theoretical accounts of how naming influences object representations and object memory. Are there naming-specific mechanisms at work, as suggested by representational shift (Lupyan, 2008, 2012)? Or are general memory principles sufficient (MacLeod et al., 2010)?

Experiment 1 combines key aspects of both labeling effect and production effect paradigms. We primarily ask whether naming of any sort, whether vocal, key-press or silent, impairs memory relative to preference rating (labeling effect), as predicted by representational shift, and whether vocal naming enhances memory relative to silent naming (production effect). Both effects may be at play, but all tasks must be tested together to observe them simultaneously. Experiment 2 specifically examines two factors that may contribute to the production effect, vocalization and uniqueness of the response.

Examining these factors may help explain differences between the naming tasks used by Lupyan (2008) versus MacLeod et al. (submitted). Experiment 3 tests additional conditions that might be responsible for the labeling effect, focusing on vocal versus key-press responses for both naming and preference rating. In Experiments 1-3, study objects were drawn from unique basic-level categories, mirroring the design used in studies of the production effect. Experiment 4 tests for both the production effect and labeling effects together using objects from only two basic-level categories, mirroring the design used in studies of the labeling effect. In all Experiments we consider whether the labeling effect or production effect is observed, and by extension whether the pattern of results is consistent with a representational shift account or a more general distinctiveness account.

EXPERIMENT 1

In Experiment 1, study objects each had unique basic-level names, following the procedure used to study the production effect. Participants performed the study tasks used in past experiments on the production effect, vocal naming and silent naming, as well as the study tasks used in past experiments on the labeling effect, key-press naming and preference rating. Old/new recognition memory was tested. By including all four study tasks in the same experiment we can gain a better understanding of the factors that produce effects of naming on memory.

For example, it is possible that vocal naming causes better memory than silent naming (production effect) but that all forms of naming cause worse memory than

preference rating (labeling effect)¹. This would suggest that both distinctiveness and representational shift are at work. Alternatively, since vocal naming is an even more overt form of object naming than key-press naming, vocal naming might cause a larger representational shift than key-press naming, causing an even larger labeling effect for vocal naming than key-press naming.

Another possibility is that impaired memory from naming observed in labeling effect studies with only two categories of objects is eliminated or even reversed when study objects come from unique basic-level categories. Indeed, Lupyan (2008) suggested that naming might actually improve recognition memory when objects are drawn from many different basic-level categories: “It is important to note that the prediction of poorer memory following overt classification holds only for *within-category* recognition. Indeed, the present account would predict that labeling study items might produce superior memory in a between-category task, because one effect of the label feedback is to ‘clean up’ the studied items to make their representations less noisy and more categorical” (p. 364). If that were true, then the nature of the lures used at test should matter. So in this experiment, half of the participants were tested with unrelated lures, requiring between-category discrimination, and the other half were tested with matched lures, requiring within-category discrimination. When there are unrelated lures from unstudied basic-level categories, the representational shift would highlight the between-category difference

¹ Predictions about silent naming vs. key-press naming are less clear-cut. On the one hand, key-press naming could lead to better recognition memory compared to silent naming. In key-press naming, participants in essence generate two responses: the name of the object, and the appropriate response key. Therefore, key-press naming may lead to better recognition than silent naming because more encoding tasks lead to more processing and thus better retention (Craik, 1979). On the other hand, Lupyan (2008) posits that overt responding with the category label induces representational shift. By this account, key-press naming should lead to worse memory than silent naming because of the more explicit response required for the key-press task.

between old and new items, improving memory for named objects. By contrast, when there are matched lures from the same basic-level category, the representational shift would make memory representations for old items less veridical, making within-category discrimination more difficult, impairing memory for named objects. So when every study object comes from a unique basic-level category, representational shift predicts a labeling effect with matched lures but no labeling effect, or even a “reverse” labeling effect, with unrelated lures. By contrast, a distinctiveness account would predict no qualitative difference for matched versus unrelated lures, only a quantitative difference because matched lures are more difficult to reject than unrelated lures.

Methods

Participants

Fifty-two members of the Vanderbilt community (24 female; mean age 24.5 years) received monetary compensation for participation. All participants were native English speakers and had normal or corrected to normal vision. Participants were randomly assigned to either the matched lure group ($n = 27$) or the unrelated lure group ($n = 26$). Data from five participants (three from the matched lure group, two from the unrelated lure group) were discarded for chance performance on the memory test (average $d' \leq 0$). Therefore, there were 24 subjects in each group in our analyses.

Stimuli

Stimuli were 240 color pictures of objects from the stimulus set created by Brady et al. (2008; <http://cvcl.mit.edu/MM/download.html>). Pictures were 256 x 256 pixels and

showed a single object on a white background. The complete stimulus set included 80 target-lure pairs (see Figure 2a) in which the target and lure were either different exemplars from the same basic-level category (exemplar pairs; top two pairs in Figure 2a) or the same object in a different state or pose (state pairs; bottom two pairs in Figure 2a). These two types of matched lures from the Brady et al. stimulus set were included to determine whether this difference would have any effect on memory, although we had no specific predictions to this effect. Note that the matched target-lure pairs used in Lupyan (2008) would be classified as exemplar pairs (see Figure 2b) within this set. Each pair of objects was from a unique basic-level category. Stimuli were randomly divided into four sets with an equal number of exemplar pairs and state pairs in each set. The object set assigned to each study task was counterbalanced across participants such that each object set was presented in a given task equally often.

For the matched lure group, one member of each pair was shown at study, and it and its matched lure were presented at test. Assignment as a target or lure was counterbalanced across participants and within each task assignment (e.g., each object in a given pair was presented as both target and lure for every task). For the unrelated lure group, one item from each pair was presented at study (counterbalanced) and lures were 80 additional pictures of objects from unique basic-level categories that were not studied.

Design

Experiment 1 used a 2 x 4 mixed design. Lure type (matched vs. unrelated) was a between-subject variable and study task (silent naming, vocal naming, key-press naming and preference rating) was a within-subjects factor. Study tasks were blocked during the

study phase, and order of targets and lures were randomized during the test phase. There were 20 target trials for each study task for each group. There were a total of 80 lures for the unrelated lure group. There were 20 matched lures for each study task for the matched lure group. Note that in the matched lure group the number of target trials were further subdivided into exemplar target/lure pairs and state target/lure pairs (10 target trials and 10 lures for each matched lure type for each study task).

Procedure

During the study phase, participants were told that they would see pictures of objects and be asked to make judgments about them. They were not explicitly told about any subsequent memory test, but they were instructed to pay careful attention and remember as much as possible about each picture. Neither the production effect nor the labeling effect has been found to depend on whether learning is intentional (MacDonald & MacLeod, 1998; Lupyan, 2008, Experiment 1-3) or incidental (MacLeod et al., 2010; Lupyan, 2008, Experiment 4). Pictures were presented for 150ms followed by a cue (question mark) for participants to respond. The cue was displayed for 2000ms regardless of how long participants took to make a response.

There were four study tasks. In the key-press naming task, participants were instructed to press the key denoting the first letter of the name of the object. In the preference task, they were instructed to rate how much they liked the object relative to other objects from the same category on a 5-point scale (1 = dislike, 5 = like). In the silent naming task, participants were instructed to silently name the object in the picture. In the vocal naming task, they were instructed to say the name of the object out loud. Tasks were

blocked², with instructions preceding each block. Each participant was assigned to one of the 24 possible task orders in each lure group. Each study object was presented once during the study phase, for a total of 80 trials (20 trials per study task). Prior to the study phase, participants were familiarized with the pace of the experiment with five practice trials. On the practice trials, participants saw pictures of chairs and lamps and were asked to press '1' for chair, and '2' for lamp.

There are some differences between the procedure used here and that used in Lupyan (2008). First, the key-press naming task had to be modified because the task used in Lupyan (2008), which was a two-alternative forced choice classification (chair vs. lamp), cannot directly accommodate dozens of study objects that all have unique names; we asked participants to press the key associated with the initial letter of the name. Second, because pilot work mirroring the conditions used by Lupyan (2008) showed that memory performance was often near ceiling when study objects came from unique basic-level categories, the following changes were implemented to increase task difficulty: shortening image exposure times during study (150ms here vs. 300ms by Lupyan), and presenting twice as many objects during study, with each object presented only once (vs. two presentations per study item by Lupyan). Finally, because of the greater complexity in response selection in the key-press naming task when all objects are unique compared to the two-alternative forced choice task used in Lupyan (2008), the following changes were made: the preference task allowed more response options (5-point "liking" scale vs. two-

² In published studies of the production effect, vocal naming and silent naming conditions have always been randomized (MacLeod et al., 2010; Ozubko & MacLeod, 2010), but the effect is also observed under blocked conditions as well (Colin MacLeod, personal communication). The labeling effect has been tested under both blocked and randomized conditions (Lupyan, 2008).

alternative “like” or “dislike”), and the response window for both preference and naming was longer (2000ms here vs. 700ms by Lupyan).

After the study phase, participants took part in another unrelated experiment (approximately 30 minutes) involving rapid categorization of gray-scale images of birds and dogs, after which they performed a recognition memory test. Regardless of the lure group they were assigned to, participants were informed that some of the pictures would be exactly the same as those they saw before, other pictures would be new but very similar to pictures they saw before, and some pictures would be brand new. Pictures were presented on the screen one at a time and participants were instructed to press ‘1’ if the picture was ‘old’ and the exact same picture they saw before, and ‘2’ if the picture was ‘new’. Pictures remained on the screen until participants made a response. No feedback was provided. There were a total of 160 test trials (80 targets and 80 matched or unrelated lures).

Results

In this and all subsequent experiments we corrected for any observed hit rate of 1.0 by subtracting .05, and for a false alarm rate of 0.0 by adding .05 (Macmillan & Kaplan, 1985). Winsorized (20%) means and variances were used in all analyses, following recommendations for robust estimation by Wilcox (2005). No qualitative effects depended on either data processing step. We report analyses of overall memory performance (sensitivity; d') in the main text. For completeness, analyses of hit and false alarm rates for each experiment are presented in Appendix A. Calculations of Cohen’s d are based on pooled standard deviations (Ellis, 2009).

The type of matched lure (exemplar vs. state) did not modulate the task effect (lure type x task interaction: $F_{3, 69} = .93$, $MSE = .20$, $p > .4$, $\eta^2_p < .04$). Therefore, we collapsed across matched lure types in our analyses³. Overall performance (d') on the recognition memory test as a function of task and lure type is plotted in Figure 3.

A significant production effect (better memory following vocal naming vs. silent naming) was observed in the unrelated lure group (mean difference = $.65 \pm .17$, $t_{23} = 8.07$, $p < .001$, Cohen's $d = 2.38$) but not the matched lure group (mean difference = $.25 \pm .30$, $t_{23} = 1.64$, $p = .115$, Cohen's $d = .48$), with a significant interaction between task and lure type ($F_{1, 46} = 5.61$, $p < .05$, $\eta^2_p = .11$).

A significant labeling effect (better memory following preference vs. key-press naming) was observed regardless of lure type (matched lures: mean difference = $.47 \pm .27$, $t_{23} = 3.64$, $p = .001$, Cohen's $d = 1.07$; unrelated lures: mean difference = $.50 \pm .18$, $t_{23} = 5.56$, $p < .01$, Cohen's $d = 1.64$; task x lure type interaction: $F_{1, 46} = .03$, $p = .87$, $\eta^2_p = .001$).

However, no labeling effect was observed when we compared vocal naming with preference rating, regardless of lure type (matched lures: mean difference = $.14 \pm .28$, $t_{23} = 1.00$, $p = .33$, Cohen's $d = .29$ unrelated lures: mean difference = $.03 \pm .14$, $t_{23} = .40$, $p = .69$, Cohen's $d = .12$; task x lure type interaction: $F_{1, 46} = .51$, $p = .48$, $\eta^2_p = .01$). Note that for both lure types the effect sizes are small, and power analyses show that to detect effects of this size would require over 300 participants.

³ Separate analyses for each type of matched lure yield the same qualitative effects as each other, and as the analyses conducted collapsed across matched lure type. Overall performance for each type of matched lure is reported in Appendix B (Table B1).

Discussion

The production effect and the labeling effect suggest potentially opposite effects of naming on visual recognition memory. For the production effect (MacLeod et al., submitted; see also MacLeod et al., 2010), vocally naming objects during study produces better memory than silently naming objects during study. For the labeling effect (Lupyan, 2008), key-press naming of objects during study produces worse memory than preference rating during study. To understand how using object names influences recognition memory, we need to systematically compare procedural variables that differ between these studies to determine what factors promote a labeling effect or production effect.

One important difference is the control task. In experiments on the production effect, vocal naming is compared to silent naming, whereas in experiments on the labeling effect, key-press naming is compared to preference rating. We had participants engage in all four study tasks. Memory following preference rating was significantly better than memory following key-press naming, replicating the labeling effect. But memory following vocal naming was comparable to memory following preference rating. Whether or not a labeling effect is observed depends on how names are used to make a response.

The representational shift account of the labeling effect (Lupyan, 2008) suggests that naming objects activates features associated with prototypical examples from the object's category, causing object representations to systematically shift towards the prototype. This representational shift causes studied items at test to appear new, causing a decrease in discriminability. It is reasonable to hypothesize that overt vocal naming activates category representations even more strongly than key-press labeling, which would cause an even greater representational shift, leading to worse memory for vocal

naming than key-press naming. That was not observed in any conditions comparing vocal naming to key-press naming.

Previous work testing the representational shift hypothesis has only tested within-category memory (Blanco & Gureckis, 2011; Lupyan, 2008; Richler et al., 2011), in which study items and lures come from two basic-level categories (e.g., chairs versus lamps). Indeed, Lupyan (2008) suggested that representational shift might in fact lead to *superior* memory following naming in a between-category task, in which study items and lures come from unique categories. Contrary to this prediction, in Experiment 1 we found that the traditional labeling effect was in fact observed when objects come from unique basic-level categories: Key-press naming produced worse recognition memory than preference rating, irrespective of whether matched or unrelated lures were used.

One possibility was that memory following both key-press naming and vocal naming would be significantly worse than memory following preference rating, indicative of a general impairment for naming. But that was not observed. Instead, whether naming impaired memory relative to preference was contingent on the modality of the naming task. A labeling effect was observed for key-press naming, but not vocal naming, for which performance was as good as preference rating. This is theoretically inconsistent with representational shift, which predicts better memory following naming when objects are from unique basic-level categories, and imposes no caveats about how naming responses are made (Lupyan, 2008). These results are also inconsistent with alternative explanations for the labeling effect (Blanco & Gureckis, 2011; Richler et al., 2011) rooted in depth of processing (e.g., rating preference is more effortful than basic-level categorization) or transfer appropriate processing (e.g., rating preference requires attention to exemplar-

specific details that are relevant to any later explicit memory test): Naming should be easier and more automatic than preference rating, or preference rating should better draw attention to features that are relevant for the memory task, regardless of how the naming response is made.

The distinctiveness account of the production effect (MacLeod et al., 2010; Ozubko & MacLeod, 2010) suggests that production improves recognition memory because only vocally named study objects are named out loud, whereas both silently named study items and new lures during test are named silently. In that way, targets and lures are better discriminated by the encoding task for items named out loud, but not items named silently. Consistent with this account, the production effect in a list discrimination task (e.g., participants must judge whether a test item was studied as part of List A or List B) is eliminated when both study and test lists contain words that are read out loud (Ozubko & MacLeod, 2010). Here we demonstrate another way that distinctiveness can be undermined. The production effect can be attenuated when matched lures are used during test. Thus, it is not only distinctiveness created by the encoding task that is critical for the production effect, but in old/new recognition tests the conceptual (and/or perceptual) distinctiveness of targets vs. lures mediates this effect as well.

It is interesting to note that Dodson and Schacter (2001) reported that production improves memory when lures are semantically related to targets by reducing rates of false memory. However, that study was based on the DRM paradigm (Deese, 1959; Roediger & McDermott, 1995) in which study targets were lists of related words (e.g., tired, bed, dream), and the critical lure was a different word that was semantically related to the set (e.g., sleep). Thus, semantic or conceptual similarity was not as strong as with our matched

lures, which were from the same basic-level category and thus shared the same name and vocal response – the metacognitive strategy of recalling having said “bell” out loud would not distinguish the target from the lure in the matched lure group in our experiment (e.g., Figure 2a). Accordingly, here false alarm rates did not differ between vocal naming and silent naming conditions (see Appendix A, Table A1).

Taken together, the results of Experiment 1 seem problematic for the representational shift account of naming effects but are largely consistent with the distinctiveness account. But why does distinctiveness improve performance following vocal naming and not key-press naming? Indeed, despite the fact that an overt response is made in the key-press naming condition but not the silent naming condition, memory following key-press naming either did not differ from (unrelated lures: mean difference = $.12 \pm .20$, $t_{23} = 1.22$, $p = .23$, Cohen’s $d = .36$) or was worse than (matched lures: mean difference = $.36 \pm .21$, $t_{23} = 3.54$, $p < .01$, Cohen’s $d = 1.04$) silent naming. According to MacLeod et al. (2010), the production effect is not simply an effect of vocalization, but requires production of an item-specific, unique response. In the key-press naming condition in Experiment 1, the response is only the first letter of the object’s name. It seems reasonable to assume that producing a single letter of a name is not as rich a production experience as vocalization, which could affect the relative distinctiveness of memory representations. In addition, the first letter of the object name could be shared by other study items⁴, decreasing distinctiveness even further for key presses compared to vocalization. Perhaps key-press naming would be a more effective study task if the

⁴ Post-hoc analysis of the object sets used in Experiment 1 found that there was only an average of 11 unique responses for the 20 objects in each object set (range: 9-12). Note that this is based on how the first author names these items, and so there may be some variability between subjects (e.g., a “parking meter” could also correctly be called a “meter”).

instructions were to type the entire name of the object, because then each object would be associated with a unique motor response sequence – production by typing rather than production by speaking. This possibility is explored in Experiment 2.

EXPERIMENT 2

The goal of Experiment 2 was to further test the distinctiveness account of the production effect and explore conditions that might increase distinctiveness in a key-press naming task. Indeed, in addition to differences in response modality (vocal vs. key-press) between studies of the production effect and labeling effect, those naming tasks also differ in the nature of the response itself. The whole name is said aloud in the vocal naming condition but only the first letter of the name is typed in the key-press condition (see Figure 1). In Experiment 2 we de-confound modality of response and uniqueness of the response by having participants respond with either the first letter of the name or the whole word either by key-press or vocally, with the aim of understanding why performance differs between the naming tasks used in studies of the production effect and labeling effect. We also included silent naming in Experiment 2 to test whether a more distinctive key-press response (typing the full name of the object) would lead to a memory advantage over silent naming, which would be consistent with a distinctiveness account, and to help compare results directly with Experiment 1. If the effect of naming on object memory can be generally explained by distinctiveness, then we should see a continuum of memory performance following the different naming tasks, with the best memory performance in the most distinctive naming condition (vocal/whole name response) – the condition used in prior studies of the production effect – and the worst memory

performance in the least distinctive naming condition (key-press/first letter) – the naming task used in prior studies of the labeling effect.

Methods

Participants

Thirty-one Vanderbilt University undergraduates (21 female; mean age 19.7 years) received course credit for participation. Data from one participant were discarded due to experimenter error. All participants were native English speakers and had normal or corrected to normal vision.

Stimuli

Stimuli were the same matched-lure pairs used Experiment 1, plus 20 additional object pairs (10 state, 10 exemplar) also obtained from the Brady et al. (2008) stimulus set.

Design

Experiment 2 used a 2 (response modality: vocal vs. key-press) x 2 (naming task: first letter vs. whole word) x 2 (matched lure type: exemplar vs. state) within-subjects design. Silent naming was also included as a control condition. Study tasks were blocked during the study phase, and targets and lures were randomized during the test phase. There were 10 target trials and 10 lure trials for each combination of response modality, naming task, and matched lure type.

Procedure

The procedure was similar to Experiment 1 with the following exceptions. First, in addition to the silent naming, key-press naming (first letter) and vocal naming (whole word) tasks used in Experiment 1, participants performed a key-press/whole word task in which they were instructed to type the full name of the object (no visual feedback was provided), and a vocal/first letter task in which they were instructed to say the first letter of the object name out loud. Thus, we crossed response modality (key-press vs. vocal) and naming task (first letter vs. whole word) in a factorial design, and included silent naming as a control condition. One hundred objects were presented at study (20 per task) and only matched lures were used in the memory test. Because five study tasks produces 120 possible task orders, task order was pseudo-randomized using a balanced Latin square. This resulted in 10 possible task orders. These were randomly assigned to participants, such that three participants completed each task order.

Results

Although matched lure type (exemplar vs. state) interacted with response modality ($F_{29} = 9.83, p < .01, \eta^2_p = .25$), this was an ordinal interaction, such that the effect of response modality was larger for state lures ($\eta^2_p = .65$) vs. exemplar lures ($\eta^2_p = .21$), but it was statistically significant in the same direction in both cases ($ps < .011$). Therefore, we collapsed across matched lure type in the subsequent analyses⁵. Overall performance (d')

⁵ Mean sensitivities for each type of matched lure are reported in Appendix B (Table B2). Separate ANOVAs for each matched lure type revealed the same qualitative effects as each other, and differ from the ANOVA on collapsed means in terms of the task effect. That is, although the main effect of task is significant when we collapse across lure type, this effect does not reach significance for either lure type when analyzed separately.

on the recognition memory test as a function of response modality and naming task is plotted in Figure 4. We conducted a 2 (response modality: key-press vs. vocal) x 2 (naming task: first letter vs. whole word) repeated-measures ANOVA, followed by planned comparisons between the key-press/first letter and vocal/whole word conditions (the two conditions that map most closely onto the tasks used in previous studies of the labeling effect and production effect, respectively, and that were used in Experiment 1), and between each key-press task (first letter, whole word) and silent naming.

Memory was better following vocal responses than key-press responses ($F_{29} = 42.23$, $p < .001$, $\eta^2_p = .59$) and when the naming task required the whole word compared with the first letter ($F_{29} = 4.44$, $p < .05$, $\eta^2_p = .13$). Accordingly, performance on the naming task used in studies of labeling effect (key-press/first letter) was significantly lower than performance on the naming task used in studies of the production effect (vocal/whole word; mean difference = $.38 \pm .10$, $t_{29} = 7.51$, $p < .001$, Cohen's $d = 1.97$). No significant interaction between response modality and naming task was observed ($F_{23} = .20$, $p = .66$, $\eta^2_p = .007$).

Replicating Experiment 1, memory performance did not differ between silent naming and key-press/first letter responses (mean difference = $.04 \pm .19$, $t_{29} = .37$, $p = .71$, Cohen's $d = .10$). More surprisingly, memory performance did not differ between silent naming and the key-press/whole word condition either (mean difference = $.12 \pm .23$, $t_{29} = 1.05$, $p = .30$, Cohen's $d = .28$). Note that for both comparisons with silent naming the effect sizes are small, and to detect effects of that size would require over 300 participants.

However, as can be appreciated from Table B2, the means are in the same direction for both lure types, such that performance is numerically better for whole word vs. key-press naming responses in all cases.

Discussion

Experiment 2 de-confounded modality of response from the number of unique responses to gain a better understanding of the factors contributing to the production effect, and the factors that lead to different levels of performance between the two naming tasks used in prior studies of the production effect and the labeling effect. Consistent with a distinctiveness account of the production effect, where memory benefits are driven by the uniqueness of responses during study (MacLeod et al., 2010), we found that more unique responses (whole word) led to better memory than responses that could be shared among multiple study items (first letter of the object name). However, a unique key-press did not produce a memory advantage over silent naming, despite the fact that no overt response is even made in the silent condition.

We also found that vocalization led to a memory benefit compared with key-press responses. Most notably, an arguably less distinct vocalization response (first letter) led to better overall memory than an arguably more distinct key-press response (whole word; mean difference = $.29 \pm .20$, $t_{29} = 3.06$, $p < .01$, Cohen's $d = .80$). One possibility is that vocalization responses of any kind are even more distinctive than key-press responses because of the auditory feedback from such a response; in addition to unique motor components of the vocalization response, they are also associated with unique auditory codes, thus adding an additional source of information to the study episode.

EXPERIMENT 3

In Experiment 2 we observed a general memory benefit for vocalization. One natural next question to ask is whether benefits from vocalization generalize to study tasks

other than naming. Indeed, we may have failed to observe a difference between vocal naming and preference rating in Experiment 1 because the preference judgment was made by key-press, not vocally. If, as in Experiment 2, the effect of vocalization is relatively additive to any other task effects, then a “labeling effect” (worse memory following naming than preference) might be observed for vocal naming if the preference response is also made vocally. To this end, in Experiment 3 we crossed the study task (naming vs. preference) with response modality (key-press vs. vocal). We were interested in whether a labeling effect (worse memory following naming compared to preference) is observed in both response modalities, and whether vocalization produces a memory benefit for both tasks.

Methods

Participants

Twenty-six Vanderbilt University undergraduates and members of the Vanderbilt University community (18 female; mean age 19.6 years) participated in the experiment in exchange for course credit or monetary compensation. Data from two participants were discarded for failing to follow task instructions (e.g., making key-press responses during a vocal response block). All participants were native English speakers and had normal or corrected to normal vision.

Stimuli

Stimuli were the same as those used in Experiment 1.

Design

Experiment 3 used a 2 (response modality: vocal vs. key-press) x 2 (study task: naming vs. preference) x 2 (matched lure type: exemplar vs. state) within-subjects design. Study tasks were blocked during the study phase, and targets and lures were randomized during the test phase. There were 10 target trials and 10 lure trials for each combination of response modality, study task, and matched lure type.

Procedure

The procedure was the same as Experiment 1 with a few exceptions: The four study tasks were key-press naming, vocal naming, key-press preference, and vocal preference. The first three tasks were identical to Experiment 1. In the vocal preference task, participants were instructed to rate how much they liked the object in the picture compared to other objects from the same category on a five-point scale, and to make their rating response out loud. Each participant was assigned to one of the 24 possible task orders. Only matched lures were presented in the recognition memory test.

Results

Because here the type of matched lure (exemplar vs. state) interacted with both response modality ($F_{1,23} = 5.19, p < .05, \eta^2_p = .18$) and task ($F_{1,23} = 13.30, p = .001, \eta^2_p = .37$) we considered the data for each matched lure type separately. The lure type x modality x task interaction was not significant ($F_{1,23} = 1.00, p = .33, \eta^2_p = .04$). Overall performance (d') on the recognition memory test as a function of matched lure type, response modality and task is plotted in Figure 5. For each lure type we conducted a 2 (response modality: vocal

vs. key-press) x 2 (task: naming vs. preference) repeated-measures ANOVA, followed by planned comparisons testing the labeling effect (naming vs. preference) for each response modality, and the production effect (benefit of vocalization) for each task.

As shown in Figure 5, for exemplar lures, a labeling effect (impaired performance for naming vs. preference) was only observed when responses were made by key-press (mean difference = $.35 \pm .20$, $t_{23} = 3.59$, $p < .01$, Cohen's $d = 1.06$) but not when they were made vocally (mean difference = $.11 \pm .28$, $t_{23} = .78$, $p = .44$, Cohen's $d = .23$); in the omnibus ANOVA, the main effects of modality ($F_{1,23} = 6.92$, $p < .05$, $\eta^2_p = .23$) and task ($F_{1,23} = 6.90$, $p < .05$, $\eta^2_p = .23$) were statistically significant, although the interaction between modality and task was not ($F_{1,23} = 2.28$, $p = .15$, $\eta^2_p = .09$). In addition, any benefit for vocalization was only observed for the naming task (mean difference = $.38 \pm .30$, $t_{23} = 2.60$, $p = .016$, Cohen's $d = .77$), not the preference task (mean difference = $.13 \pm .22$, $t_{23} = 1.29$, $p = .21$, Cohen's $d = .38$).

For state lures, memory was better following vocal responses than key-press responses ($F_{1,23} = 55.70$, $p < .001$, $\eta^2_p = .71$). Neither the main effect of task ($F_{1,23} = 1.34$, $p = .26$, $\eta^2_p = .06$) nor the interaction were significant ($F_{1,23} = .08$, $p = .78$, $\eta^2_p = .004$). No labeling effect was observed for either response modality (key-press: mean difference = $.08 \pm .30$, $t_{23} = .54$, $p = .60$, Cohen's $d = .16$; vocal: mean difference = $.13 \pm .21$, $t_{23} = 1.25$, $p = .22$, Cohen's $d = .39$). Note that sample sizes of more than 300 participants would be required to observe effects of this size. A benefit of vocalization was observed for both tasks (naming: mean difference = $.55 \pm .26$, $t_{23} = 4.31$, $p < .001$, Cohen's $d = 1.27$; preference: mean difference = $.50 \pm .20$, $t_{23} = 5.17$, $p < .001$, Cohen's $d = 1.52$).

Discussion

In Experiment 1, we observed no labeling effect (impaired memory following naming compared to preference) when the naming response was made vocally. However, in that case, vocal naming was compared to a preference rating made by key-press. Perhaps if both preference and naming responses were made vocally, we might see the labeling effect reemerge. To test this, in Experiment 3 task (naming versus preference) and response modality (vocal versus key-press) were manipulated factorially; following Experiments 1 and 2, we also included both exemplar and state change matched lures at test.

Naming only led to significantly worse memory than preference rating when both responses were made by key-press and when the lures were different exemplars, the conditions that most closely matched to the original Lupyan (2008) study. Neither with vocal responses nor with state change lures was a labeling effect observed. Note that the labeling effect was significant for state change lures in Experiment 1 ($t_{23} = 2.55, p = .018$), and it is unclear why we did not replicate that result in Experiment 3. Overall, these results seem difficult to reconcile with the representational shift hypothesis. Naming objects at study should modify stored object representations, regardless of whether naming responses are made vocally or by key-press. These results are also difficult to explain by an alternative account proposed by Richler et al. (2011), who suggested that rating preference might be more effortful than naming and require attention to exemplar-specific details that would improve later memory, consistent with depth of processing or transfer appropriate processing. If true, then vocal preference ratings should lead to better memory than vocal

naming, which was not observed. It appears that any benefits of vocalizing or rating preference are not simply additive.

There was generally an overall benefit of vocalizing compared to key-press responses. Memory was significantly better for vocalized responses than key-press responses, except for preference rating in the exemplar lure condition. The better memory following vocal versus key-press naming was anticipated based on the results of Experiment 1. Better memory following vocal versus key-press preference rating was also observed in the state change condition. So even in a condition in which many objects were given the same preference rating, a vocal response produced better memory than a key-press response. Perhaps subtle variability in how those similar preference ratings are made vocally is sufficient to create more distinctiveness to lead to better memory.

EXPERIMENT 4

So far we have tested the labeling and production effects in conditions in which each study object is from a unique basic-level category, similar to prior studies of the production effect. In Experiment 4, study objects were drawn from only two categories (chairs and lamps), matching the procedure used in studies of the labeling effect (Lupyan, 2008; Richler et al., 2011). We expect to replicate the labeling effect, with lower memory following key-press naming than preference rating. What about vocal naming? According to the distinctiveness account of the production effect, vocal naming helps provide a signature for discriminating study items from lures. But in the case of only two categories, the same two signatures (“chair” or “lamp”) apply to all study items. Therefore, distinctiveness would predict a significantly attenuated, and possibly eliminated, production effect in this

experiment. As in Experiment 1, we are interested in whether there is a labeling effect, whether the labeling effect depends on the modality of the naming task, and whether there is a production effect.

Methods

Participants

Fifty-one Vanderbilt University undergraduates (37 female; mean age 19.2 years) participated in the experiment in exchange for course credit. All participants were native English speakers and had normal or corrected to normal vision. Participants were randomly assigned to either the matched lure group ($n = 26$) or the unrelated lure group ($n = 25$). Data from three participants (two from the matched lure group, one from the unrelated lures group) were discarded for below chance performance on the memory test (average $d' \leq 0$). Thus, there were 24 participants in each lure group in the analyses.

Stimuli

Stimuli were 240 pictures of chairs and lamps (120 from each category) downloaded from the Ikea website (www.ikea.com). Pictures were 250 x 250 pixels and showed a single object on a white background. The complete stimulus set included 80 target-lure pairs like those used in Lupyan (2008) and Richler et al. (2011), in which the target and lure were very similar exemplars that differed in subtle details, such as the presence or absence of a single feature (see Figure 2b for examples). Stimuli were randomly divided into four sets with an equal number of chairs and lamps in each set. The object set assigned to each study task was counterbalanced across participants.

For the matched lure group, one item from each pair was shown at study, and the matched lure was presented at test. Assignment as a target or lure was counterbalanced across participants. For the unrelated lure group, one item from each pair was presented at study (counterbalanced) and lures were 80 additional chair and lamp pictures that were not studied and that were not obviously matched to other study items. Note that unlike Experiment 1 unrelated lures still shared the same labels (chair or lamp) as targets. In this sense, the unrelated lures differed from targets in terms of visual similarity only, whereas unrelated lures also differed in conceptual similarity in Experiment 1.

Design

Experiment 4 used a 2 x 4 mixed design. Lure type (matched vs. unrelated) was a between-subjects factor, and study task (silent naming, vocal naming, key-press naming and rating preference) was a within-subjects factor. Study task was blocked during the study phase, and targets and lures were randomized in the test phase. There were 20 target trials for each study task.

Procedure

The study phase was similar to Experiment 1, with the following exceptions. In Experiment 4 the response instructions for the key-press naming and preference tasks were identical to the tasks used by Lupyan (2008) and Richler et al. (2011). In the key-press task, participants were instructed to press '1' if the object was a chair, and '2' if the object was a lamp. In the preference task, participants pressed '1' if they liked the object and '2' if they disliked the object. Also, like Lupyan (2008) and Richler et al. (2011), in

Experiment 4 images were presented for 300ms (rather than 150ms, as in Experiments 1-3). Each participant was assigned to one of the 24 possible task orders in each lure group.

Prior to the study phase, participants were familiarized with the pace of the task with five practice trials. On the practice trials, participants saw pictures of tables and were asked to press '1' if the table was round, and '2' if the table was square.

The test phase was identical to Experiment 1, with the exception that there was no intervening distractor task between study and test phases.

Results

Overall performance (d') on the recognition memory test as a function of lure type and task is plotted in Figure 6.

The labeling effect (impaired memory following key-press naming vs. preference rating) was observed for both lure groups (matched lures: mean difference = $.65 \pm .17$, $t_{23} = 7.57$, $p < .001$, Cohen's $d = 2.23$; unrelated lures: mean difference = $1.04 \pm .12$, $t_{23} = 17.59$, $p < .001$, Cohen's $d = 5.19$), and the effect was larger for unrelated lures ($F_{1,46} = 13.84$, $p = .001$, $\eta^2_p = .23$). In contrast to Experiment 1, a labeling effect was observed regardless of that modality of the naming task, with significantly worse memory following vocal naming relative to preference rating (matched lures: mean difference = $.74 \pm .22$, $t_{23} = 7.00$, $p < .001$, Cohen's $d = 2.06$; unrelated lures: mean difference = $.96 \pm .13$, $t_{23} = 14.81$, $p < .001$, Cohen's $d = 4.36$; task x lure type interaction: $F_{1,46} = 3.27$, $p = .08$, $\eta^2_p = .07$).

Like Experiment 1, we did observe a small production effect (better memory following vocal naming vs. silent naming) for the unrelated lure group (mean difference = $.18 \pm .17$, $t_{23} = 2.22$, $p < .05$, Cohen's $d = .65$) but not the matched lure group (mean

difference = $.05 \pm .17$, $t_{23} = .64$, $p = .53$, Cohen's $d = .19$), however, the interaction between task and lure type was not significant ($F_{1, 46} = 1.25$, $p = .27$, $\eta^2_p = .03$).

Discussion

Experiment 4 used only two categories of objects (chairs and lamps), adopting the design of prior experiments on the labeling effect. Not surprisingly, there was a category length effect, and overall performance was much lower in this experiment compared to Experiments 1-3 in which there was only one exemplar per category at study (see Malmberg et al., 2012; Neely & Tse, 2009). More critically, unlike Experiment 1, for both key-press and vocal naming, we observed significant labeling effects (worse memory following naming than preference). Any production effect (better memory following vocal naming) was significantly attenuated, only exceeding the criterion for statistical significance in the unrelated lure condition.

Although it is possible that representational shift (Lupyan, 2008) is sensitive to task context and only occurs for both vocal and key-press naming when there are only a small number of object categories, a more parsimonious explanation seems to be that any advantage for vocal naming, like that observed in the previous experiments, requires that object names be unique, consistent with the distinctiveness account (MacLeod et al., 2010).

GENERAL DISCUSSION

What effect does naming objects at study have on visual recognition memory for objects at test? Lupyan's account for the labeling effect suggests that naming impairs memory because objects studied by key-press naming are recognized worse than objects

studied by preference rating (Lupyan, 2008). By contrast, MacLeod et al.' account of the production effect suggests that naming enhances memory because objects studied by vocal naming are recognized better than objects studied by silent naming (MacLeod, 2010; MacLeod et al., submitted). In this article, we systematically compared various experimental factors that differ between studies of the labeling effect and the production effect (see Figure 1) to better understand how naming object at study influences recognition memory for those objects at test. We consider what these results suggest theoretically about the potential impact of naming on memory representations, such as whether they are best explained by assuming special naming-specific mechanisms or more general memory mechanisms. A summary of the general experimental variables manipulated across the four experiments is provided in Table 1.

Table 1. Summary of main experimental variables manipulated in Experiments 1-4.

Experiment	Lure Type(s)	# Studied Categories	Study Task	Study Modality	Study Response
1	matched or semantically unrelated	Many	naming	silent	whole word
			naming	vocal	whole word
			naming	key-press	first letter
			preference	key-press	1-5 scale
2	matched	Many	naming	silent	whole word
			naming	vocal	first letter
			naming	vocal	whole word
			naming	key-press	first letter
			naming	key-press	whole word
3	matched	Many	naming	vocal	whole word
			preference	vocal	1-5 scale
			naming	key-press	first letter
			preference	key-press	1-5 scale
4	matched or physically unrelated	Two	naming	silent	whole word
			naming	vocal	whole word
			naming	key-press	2-AFC
			preference	key-press	2-AFC

According to the representational shift account (Lupyan, 2008), naming an object causes a systematic change in its representation, making memory for that object less veridical than it would be otherwise. In Experiments 1 and 3, we replicated the standard labeling effect, whereby key-press naming at study led to worse memory than preference rating at test. Ironically, this replication of the labeling effect seems to provide evidence against representational shift. When studied objects come from many different categories, as in Experiments 1 and 3, representational shift predicts that naming should actually *improve* memory compared to preference rating (Lupyan, 2008). It did not. Whether matched or unrelated lures were used at test, objects studied by key-press naming produced worse recognition memory than objects studied by preference rating.

Furthermore, if representational shift occurs whenever an object is named, then vocal naming should produce at least as much of a representational shift as key-press naming, if not more. Yet, in Experiments 1 and 3, vocally naming negated the labeling effect entirely. Recognition memory after vocal naming at study was the same as recognition memory after preference rating at study.

Although these findings appear incompatible with the representational shift account of the labeling effect, they appear to support and extend the distinctiveness account of the production effect (MacLeod et al., 2010; Ozubko & MacLeod, 2010). In Experiment 2 we observed that increasing the uniqueness of responses during study leads to better memory, and in Experiment 4 we observed that the production effect is significantly diminished when all study objects share only one of two possible category names. These results suggest that the benefit for vocal naming is contingent on having item-specific, unique responses because then the memory representations created during encoding task help to distinguish

between the studied items and new items later at test (MacLeod et al., 2010). Indeed, it may be that any task that produces unique responses for each study item, regardless of whether the response involves names or not, could be as effective as vocal naming. MacLeod et al. (2010) showed that a production effect for visually presented words does not require overt vocal production (the effect is obtained when words are mouthed silently), nor does it require that the response be semantically meaningful (the effect is obtained when non-words are read out loud as well). In this sense, the production effect when naming objects seems to be a manifestation of more general memory phenomena.

We also found that the production effect was attenuated when targets and lures shared the same names (matched lure condition in Experiment 1), suggesting that it is not sufficient for the encoding task to distinguish between targets and lures (Ozubko & MacLeod, 2010), but targets and lures must also be distinct conceptually. Visual distinctiveness may also be involved because objects from the same category are both conceptually related and visually similar, but because naming responses highlight conceptual information as well, we suspect that this plays a larger role in mediating the production effect.

Across experiments, preference rating led to consistently good memory. In all experiments, preference rating led to better memory than key-press naming. In Experiments 1 and 3, preference rating led to comparable memory as vocal naming, even when the name was diagnostic of whether the item was old or new (Experiment 1, unrelated lure condition), and even though preference ratings themselves are not as unique as naming responses (in that many different objects had to be given the same preference rating). There are several reasons why preference rating might be a particularly effective

study task. Preference ratings might focus attention on exemplar-specific details (Richler et al., 2011) that are relevant for subsequent memory – a participant might think, “I like *this* shoe because of the buckles”. While multiple items may share the same preference rating, they may each be preferred for different reasons. Or preference ratings might produce a self-reference effect (Symons & Johnson, 1997) – a participant is ultimately rating how *they* feel about a given object. Interestingly, there is evidence that the benefit of vocal naming may also be related to self-reference: The production effect is reduced when the participant makes a vocal response at the same time as another person (MacLeod, 2011). However, an even simpler possibility is that, like the advantage for vocal naming, the memory advantage for preference rating is due to distinctiveness: Rating preference may lead to more distinctive encoding episodes due to the affective nature of the response (Lebrecht & Tarr, 2010) or because it is a more effortful judgment (Richler et al., 2011).

Of course, these theories about the preference task are purely speculative. Moreover, preference does not universally improve memory compared to naming because vocal naming and preference (whether vocal or key-press) can produce comparable memory (Experiment 3). We do not have a simple *a priori* explanation for why key-press naming (regardless of whether the response is the first letter of the name or the whole word) seems to be a particularly poor study task, resulting in performance that is no better than a silent naming control (Experiment 2) despite the fact that overt responses are made.

Our ultimate goal in this series of experiments was not to uncover the precise mechanisms underlying each possible study task variant that we used across the different experiments. Rather, these task manipulations were in the service of a theoretical question: How does using names influence object memory? Across four experiments we found that

three different tasks that required naming of some sort (silent naming, key-press naming, and vocal naming) failed to produce equivalent memory across experimental contexts. Not only did performance differ between naming tasks, but these differences also depended on other aspects of the experiment, such as lure type and the number of basic-level categories. Although this is not surprising in the case of the production effect, given the emphasis on distinctiveness as the underlying mechanism (MacLeod et al., 2010; Ozubko & MacLeod, 2010), this is problematic for the representational shift account, which depends on the explicit use of category labels (Lupyan, 2008, 2012). Thus, our results suggest that “naming” does not exert a special influence on memory (Lupyan, 2008, 2012), as its effects can be explained in terms of more general memory phenomena such as distinctiveness (MacLeod et al., 2010) or the type of processing it entails (Richler et al., 2011).

REFERENCES

- Ashby, F.G., Alfonso-Reese, L.A., Turken, A.U., & Waldron, E.M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Bulletin, 105*, 442-481.
- Blanco, N. and Gureckis, T.M. (2011). *Does category labeling lead to forgetting?* In L. Carlson, C. Höscher and T. Shipley (Eds), Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, TX: Cognitive Science Society
- Brady, T.F., Konkle, T., Alvarez, G.A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences, 105*, 14325-14329.
- Coltheart, V. (1999). Comparing short-term memory and memory for rapidly presented visual stimuli. *International Journal of Psychology, 34*, 293-300.
- Conway, M.A., & Gathercole, S.E. (1987). Modality and long-term memory. *Journal of Memory and Language, 26*, 341-361.
- Craik, F.I.M. (1979). Human memory. *Annual Reviews Psychology, 30*, 63-102.
- Craik, F.I.M., & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*, 671-684.
- Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*, 17-22.
- 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*, 155-161.
- Ellis, P.D. (2009). "Effect size calculators",
<http://www.polyu.edu.hk/mm/effectsizefaqs/calculator/calculator.html>.

- Goldstone, R. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*, 178-200.
- Joyce, C. & Cottrell, G.W. (2004). Solving the visual expertise mystery. In *Connectionist Models of Cognition and Perception II: Proceedings of the Eighth Neural Computation and Psychology Workshop*, Howard Bowman and Christophe Labiouse (Eds.), World Scientific.
- Kruschke, J.K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.
- Lamberts, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, *107*, 227-260.
- Lebrecht, S., & Tarr, M.J. (2010). Object recognition: Features and categories defining an object's micro-valence through implicit measures. *Journal of Vision*, *10*, 966.
- Lupyan, G. (2008). From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*, *137*, 348-369.
- Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. *Frontiers in Psychology*, *3*, 1-13.
- MacDonald, P.A., & MacLeod, C.M. (1998). The influence of attention at encoding on direct and indirect remembering. *Acta Psychologica*, *98*, 291-310.
- MacLeod, C.M. (2010). When learning met memory. *Canadian Journal of Experimental Psychology*, *64*, 227-240.
- MacLeod, C.M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin & Review*, *18*, 1197-1202.

- 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*, *35*, 671-685.
- MacLeod, C.M., Ozubko, J.D., Forrin, N.D., & Hourihan, K.L. (submitted). Assessing the basis of the production effect: Evaluating the distinctiveness account.
- Macmillan, N.A., & Kaplan, H.L. (1985). Detection theory analysis of group data: Estimating sensitivity from average hit and false alarm rates. *Psychological Bulletin*, *98*, 185-199.
- Malmberg, K.J., Criss, A.H., Gangwani, T.H., & Shiffrin, R.M. (2012). Overcoming the negative consequences of interference from recognition memory testing. *Psychological Science*, *23*, 115-119.
- Meyer, A., & Damian, M. (2007). Activation of distractor names in the picture-picture interference paradigm. *Memory & Cognition*, *35*, 494-503.
- Morsella, E., & Miozzo, M. (2002). Evidence for a cascade model of lexical access in speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *28*, 555-563.
- Navarette, E., & Costa, A. (2005). Phonological activation of ignored pictures: Further evidence for a cascade model of lexical access. *Journal of Memory and Language*, *53*, 359-377.
- Neely, J.H., & Tse, C.-S. (2009). Category length produces an inverted-U discriminability function in episodic recognition memory. *The Quarterly Journal of Experimental Psychology*, *62*, 1141-1172.

- Nosofsky, R.M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 700-708.
- Nosofsky, R.M., & Palmeri, T.J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review*, *104*, 266-300.
- 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.
- Palmeri, T.J., & Tarr, M.J. (2008). Object recognition and long-term visual memory for objects. In S. Luck & A. Hollingsworth (Eds.), *Visual Memory*. Oxford University Press.
- Richler, J.J., Gauthier, I., & Palmeri, T.J. (2011). Basic-level categorization accounts for naming effects in visual recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *37*, 1579-1587.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*, 1019-1025.
- Roberson, D., & Davidoff, J. (2000). The categorical perception of colors and facial expressions: The effect of verbal interference. *Memory & Cognition*, *28*, 977-986.
- Roediger, H.L., III, & McDermott, K.B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 803-814.
- Schiano, D.J., & Watkins, M.J. (1981). Speech-like coding of pictures in short-term memory. *Memory & Cognition*, *9*, 110-114.

- Seamon, J.G., Luo, C.R., Schlegal, S.E., Greene, S.E., & Goldenberg, A.B. (2000). False memory for categorized pictures and words: The category associates procedure for studying memory errors in children and adults. *Journal of Memory and Language, 42*, 120-146.
- Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Science, 104*, 6424-6429.
- Symons, C.S., & Johnson, B.T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin, 121*, 371-394.
- Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Academic Press.
- Zelinsky, G.J., & Murphy, G.L. (2000). Synchronizing visual and language processing: An effect of object name length on eye movements. *Psychological Science, 11*, 125-131.

APPENDIX A: Analyses of Hit and False Alarm Rates for Experiments 1-4

Experiment 1

Table A1. Mean hit and false alarm rates for each condition in Experiment 1 (standard errors in brackets). In the matched lure condition different lures are associated with targets studied in different encoding conditions; in the unrelated lure condition the lures are not condition-specific.

Lure Type	Task	Hit Rate	False Alarm Rate
Matched	Silent Naming	.75 (.02)	.23 (.02)
	Vocal Naming	.84 (.02)	.26 (.02)
	Key-press Naming	.71 (.02)	.29 (.02)
	Preference Rating	.80 (.02)	.25 (.02)
Unrelated	Silent Naming	.76 (.02)	
	Vocal Naming	.91 (.01)	
	Key-press Naming	.81 (.02)	
	Preference Rating	.91 (.01)	
			.05 (.01)

There was a significant production effect in hit rate for both lure groups (matched lures: mean difference = $.09 \pm .04$, $t_{23} = 4.22$, $p < .001$, Cohen's $d = 1.24$; unrelated lures: mean difference = $.16 \pm .04$, $t_{23} = 7.95$, $p < .001$, Cohen's $d = 2.34$), but the effect was significantly larger for the unrelated lure group ($F_{1,46} = 4.81$, $p < .05$, $\eta^2_p = .10$). There was also a significant labeling effect for both lure groups (matched lures: mean difference = $.09 \pm .05$, $t_{23} = 3.93$, $p = .001$, Cohen's $d = 1.16$; unrelated lures: mean difference = $.10 \pm .04$, $t_{23} = 6.12$, $p < .001$, Cohen's $d = 1.80$; task x lure type interaction: $F_{1,46} = .20$, $p = .66$, $\eta^2_p = .004$). Critically, there was no labeling effect when we compared vocal naming and preference rating for either lure group (matched lures: mean difference = $.04 \pm .05$, $t_{23} = 1.72$, $p = .099$, Cohen's $d = .51$; unrelated lures: mean difference = $.01 \pm .02$, $t_{23} = .617$, $p = .543$, Cohen's $d = .18$; task x lure type interaction: $F_{1,46} = 1.84$, $p = .18$, $\eta^2_p = .04$).

In the matched lure group, no significant difference in false alarms were observed between vocal and silent naming (mean difference = $.03 \pm .05$, $t_{23} = 1.33$, $p = .195$, Cohen's $d = .39$). There was a significant difference in false alarms observed following key-press naming vs. preference rating (mean difference = $.05 \pm .04$, $t_{23} = 2.53$, $p = .019$, Cohen's $d = .75$).

Experiment 2

Table A2. Mean hit and false alarm rates for each condition in Experiment 2 (standard error shown in brackets).

Response Modality	Naming Task	Hit Rate	False Alarm Rate
	Silent	.69 (.03)	.21 (.01)
Key Press	First Letter	.71 (.02)	.22 (.01)
	Whole Word	.76 (.01)	.24 (.02)
Vocal	First Letter	.79 (.02)	.18 (.02)
	Whole Word	.84 (.01)	.18 (.01)

In hit rate there was a main effect of response modality ($F_{29} = 61.61$, $p < .001$, $\eta^2_p = .68$), a main effect of naming task ($F_{29} = 15.33$, $p < .001$, $\eta^2_p = .35$) but no interaction ($F_{29} = .003$, $p = .96$, $\eta^2_p < .001$). Whereas silent naming did not differ from the key-press/first letter condition (mean difference = $.02 \pm .05$, $t = .82$, $p = .42$, Cohen's $d = .22$), memory was better in the key-press/whole word condition compared with silent naming (mean difference = $.07 \pm .06$, $t = 2.24$, $p < .05$, Cohen's $d = .59$).

There were more false alarms for the key-press condition than for the vocal responses ($F_{29} = 27.77$, $p < .001$, $\eta^2_p = .49$). Neither the main effect of naming task nor the interaction between response modality and naming task were significant ($ps > .2$, $\eta^2_p < .05$). False alarm rates did not differ between silent naming and either of the key-press conditions ($ps > .2$, Cohen's $ds < .15$).

Experiment 3

Table A3. Mean hit and false alarm rates for each condition in Experiment 3 (standard error shown in brackets).

Lure Type	Response Modality	Task	Hit Rate	False Alarm Rate
Exemplar	Vocal	Naming	.83 (.02)	.17 (.02)
		Preference	.80 (.02)	.11 (.01)
	Key-Press	Naming	.80 (.02)	.20 (.02)
		Preference	.81 (.02)	.15 (.01)
State	Vocal	Naming	.88 (.01)	.25 (.03)
		Preference	.86 (.01)	.24 (.02)
	Key-Press	Naming	.83 (.01)	.34 (.02)
		Preference	.81 (.02)	.35 (.03)

For exemplar lures there were no significant main effects or interactions in hit rate ($ps > .2$, $\eta^2_p < .08$). For state lures, hit rate was higher following vocal responses ($F_{23} = 20.80$, $p < .001$, $\eta^2_p = .48$). Neither the main effect of task nor the interaction between response modality and task were significant ($ps > .2$, $\eta^2_p < .07$).

For exemplar lures, there was a trend toward more false alarms following key-press responses ($F_{23} = 3.90$, $p = .06$, $\eta^2_p = .15$), and significantly more false alarms following naming vs. rating preference ($F_{23} = 14.63$, $p = .001$, $\eta^2_p = .39$). The interaction between response modality and task was not significant ($F_{23} = .01$, $p = .76$, $\eta^2_p = .004$). For state lures, there were more false alarms following key-press responses ($F_{23} = 16.87$, $p < .001$, $\eta^2_p = .42$). Neither the main effect of task nor the interaction were significant ($ps > .6$, $\eta^2_p < .02$).

Experiment 4

Table A4. Mean hit and false alarm rates for each condition in Experiment 4 (standard errors in brackets). Note that in the matched lure condition different lures are associated with targets studied in different encoding conditions; in the unrelated lure condition, lures are not condition-specific.

Lure Type	Task	Hit Rate	False Alarm Rate
Matched	Silent Naming	.47 (.02)	.28 (.02)
	Vocal Naming	.48 (.02)	.30 (.02)
	Key-press Naming	.50 (.02)	.27 (.02)
	Preference Rating	.84 (.02)	.41 (.02)
Unrelated	Silent Naming	.50 (.02)	
	Vocal Naming	.55 (.02)	
	Key-press Naming	.54 (.03)	
	Preference Rating	.84 (.02)	
			.19 (.02)

A labeling effect was observed for both lure groups in hit rates (matched lures: mean difference = $.34 \pm .04$, $t_{23} = 18.26$, $p < .001$, Cohen's $d = 5.38$; unrelated lures: mean difference = $.30 \pm .06$, $t_{23} = 10.18$, $p < .001$, Cohen's $d = 3.00$; task x lure type interaction: $F_{1,46} = 1.74$, $p = .19$, $\eta^2_p = .04$). Worse memory for naming relative to preference was also observed for the vocal naming task in both lure groups (matched lures: mean difference = $.36 \pm .05$, $t_{23} = 15.62$, $p < .001$, Cohen's $d = 3.00$; unrelated lures: mean difference = $.29 \pm .07$, $t_{23} = 9.56$, $p < .001$, Cohen's $d = 2.81$), and the effect was larger for the matched lure group ($F_{1,46} = 4.15$, $p < .05$, $\eta^2_p = .08$). There was no production effect in hit rates for the matched lure group (mean difference = $.01 \pm .05$, $t_{23} = .40$, $p = .70$, Cohen's $d = .12$) and a marginally significant production effect in hit rates for the unrelated lure group (mean difference = $.05 \pm .07$, $t_{23} = 2.01$, $p = .056$, Cohen's $d = .59$), but the interaction between task and lure type was not significant ($F_{1,46} = 1.36$, $p = .25$, $\eta^2_p = .03$).

In the matched lure group, the production effect tasks did not lead to differences in false alarms (mean difference = $.01 \pm .04$, $t_{23} = .79$, $p = .44$, Cohen's $d = .23$). There was a

significant difference in false alarms for the two conditions in the labeling effect, with more false alarms following preference rating vs. key-press naming (mean difference = $.14 \pm .05$, $t_{23} = 5.14$, $p < .001$, Cohen's $d = 1.52$).

APPENDIX B

Table B1. Overall memory performance (d') for each condition in Experiment 1 for each type of matched lure (standard error in brackets).

Matched Lure Type	Task	Sensitivity (d')
Exemplar	Silent Naming	1.64 (.08)
	Vocal Naming	1.66 (.14)
	Key-press Naming	1.22 (.13)
	Preference Rating	1.72 (.13)
State	Silent Naming	1.41 (.11)
	Vocal Naming	1.68 (.14)
	Key-press Naming	1.09 (.11)
	Preference Rating	1.49 (.11)

Table B2. Overall memory performance (d') for each condition in Experiment 2 for each type of matched lure (standard error in brackets).

Matched Lure Type	Response Modality	Naming Task	Sensitivity (d')
Exemplar	Silent		1.49 (.12)
		Key Press	1.66 (.08)
	Vocal	First Letter	1.74 (.11)
		Whole Word	1.86 (.15)
		First Letter	1.98 (.08)
		Whole World	1.98 (.08)
State	Silent		1.31 (.12)
		Key Press	1.28 (.11)
	Vocal	First Letter	1.39 (.05)
		Whole Word	1.80 (.10)
		First Letter	2.00 (.09)
		Whole World	2.00 (.09)

ACKNOWLEDGEMENTS

This work was funded by the Temporal Dynamics of Learning Center (SBE-0542013), an NSF Science of Learning Center, and a grant from the James S. McDonnell Foundation. We would like to thank Justin Barisich, Steph Harrison, Kaleb Lowe and Lisa Weinberg for assistance with data collection.

FIGURE CAPTIONS

Figure 1. An illustration of three of the experimental factors that differ between experiments on the production effect and experiments on the labeling effect. Also illustrated are idealized data for the production effect (better memory following vocal naming vs. silent naming) and the labeling effect (worse memory following key-press naming vs. preference rating).

Figure 2. Examples of target-lure pairs used in Experiments 1-3 (a) and Experiment 4 (b). In panel A the top two examples show exemplar-pairs, and the bottom two examples show state-pairs. The last pair of lamps in Panel B actually differed in color.

Figure 3. Overall memory performance (d') in Experiment 1 as a function of study task and lure group. Error bars show 95% confidence intervals for the critical paired t-tests (production effect: silent naming vs. vocal naming; labeling effect: key-press naming vs. rating preference).

Figure 4. Overall recognition memory performance (d') in Experiment 2 as a function of response modality and naming task. Error bars show 95% confidence intervals of the within-subject effect of "Task" based on a repeated-measures ANOVA that included all five conditions.

Figure 5. Overall memory performance (d') in Experiment 3 as a function of response modality and task for each type of matched lure. Error bars show 95% confidence intervals for the labeling effect (naming vs. preference) in each condition.

Figure 6. Overall memory performance (d') in Experiment 4 as a function of study task and lure type. Error bars show 95% confidence intervals for the critical paired t-tests (production effect: silent naming vs. vocal naming; labeling effect: key-press naming vs. rating preference).

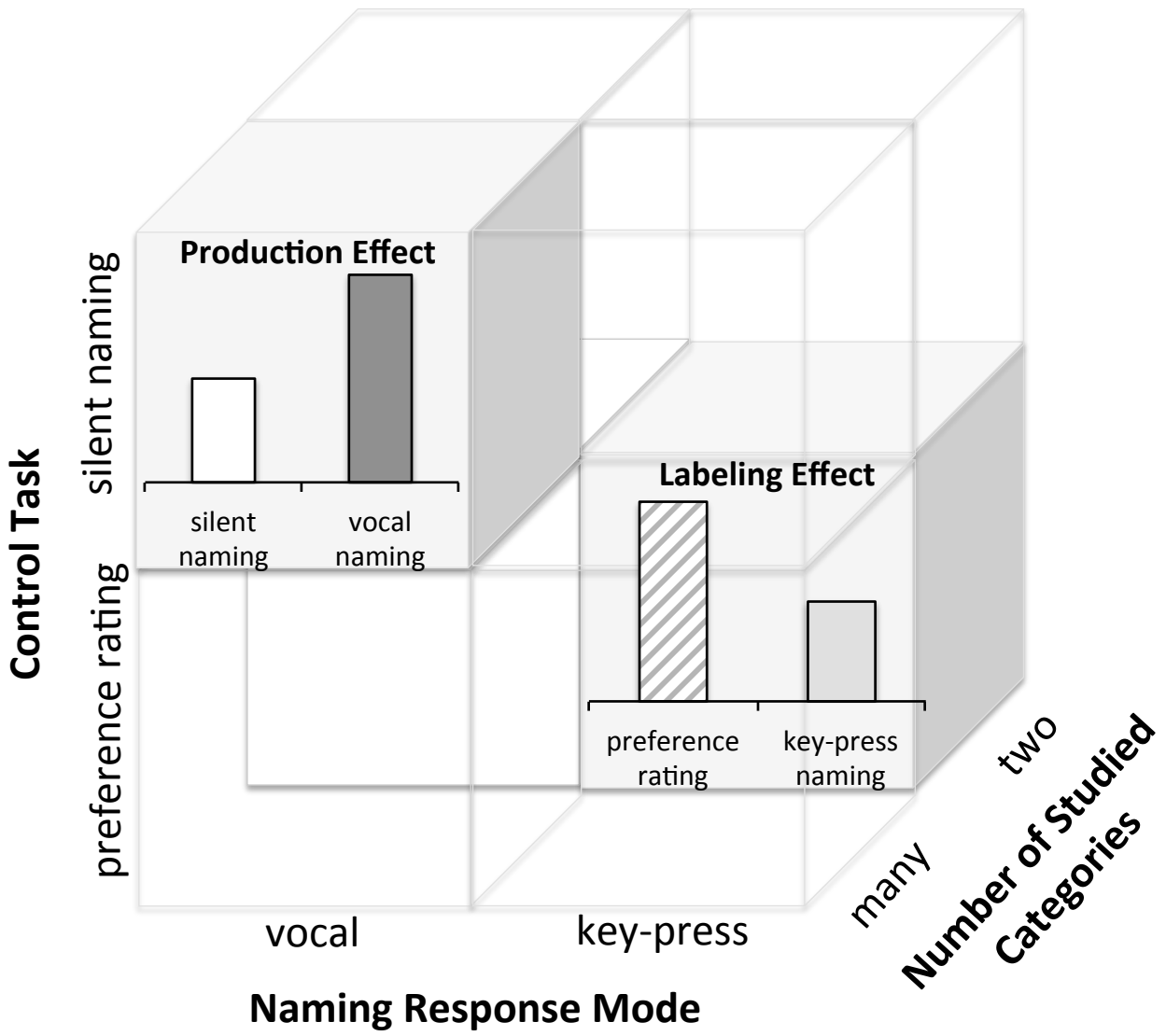


Figure 1

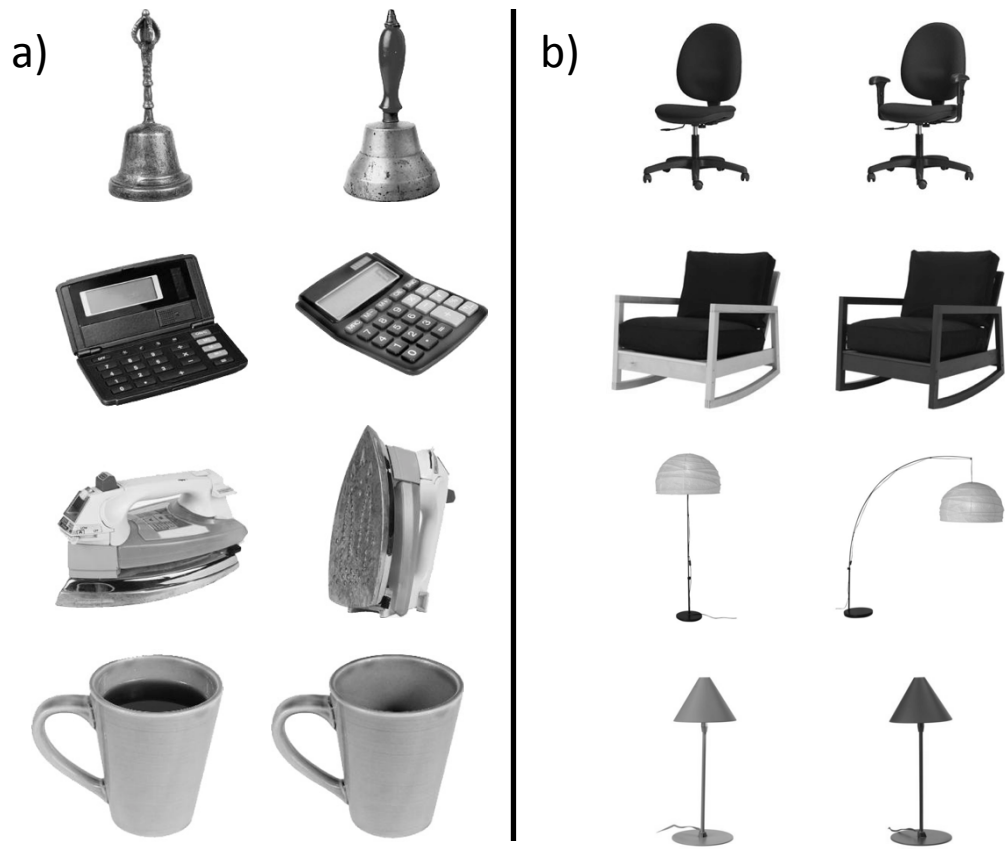


Figure 2

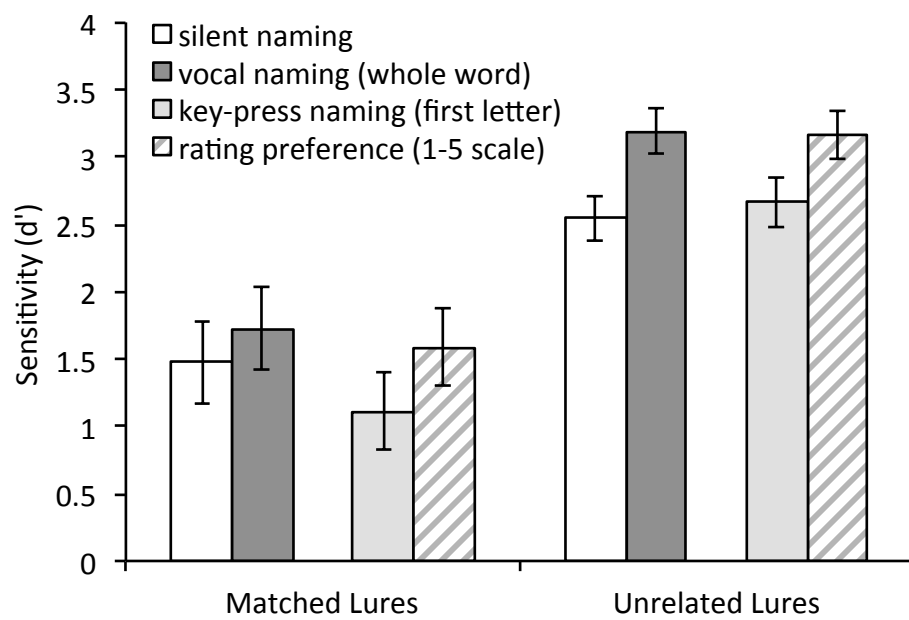


Figure 3

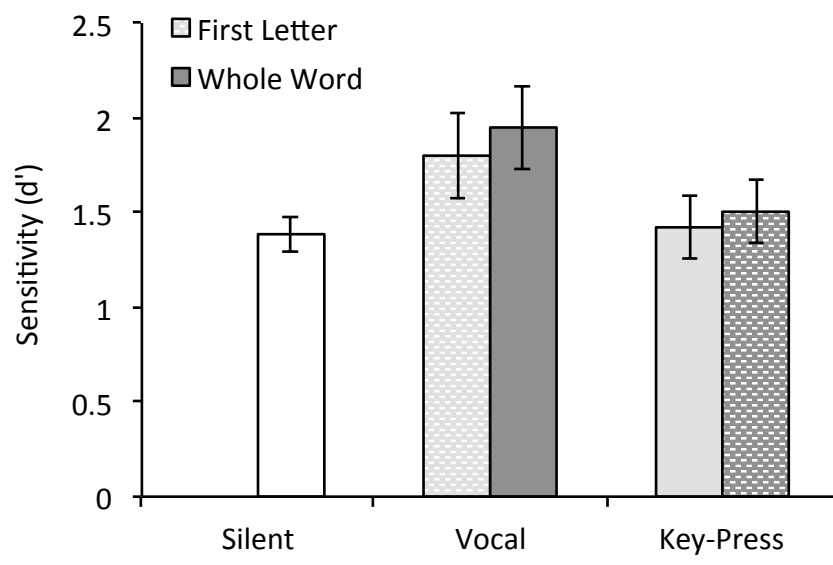


Figure 4

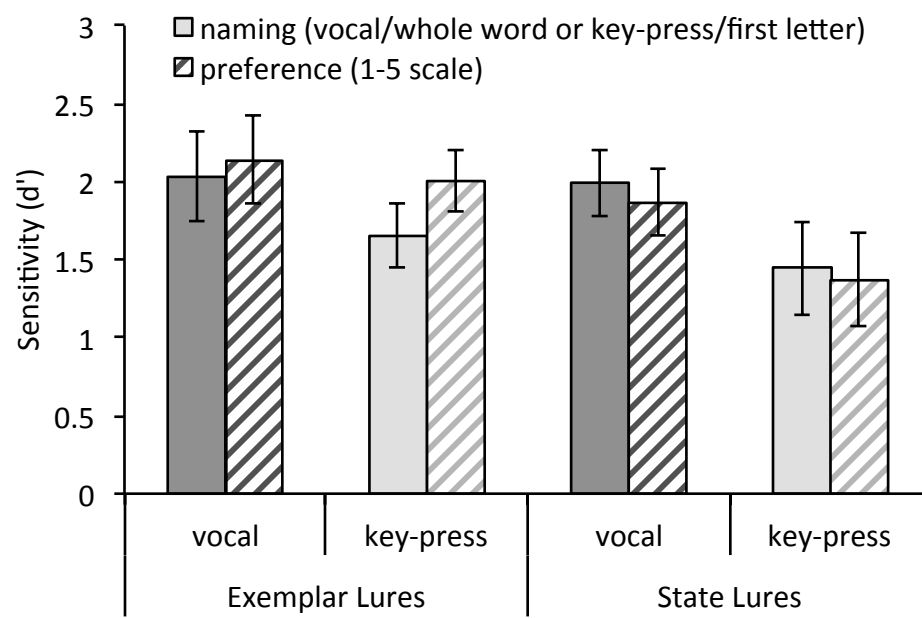


Figure 5

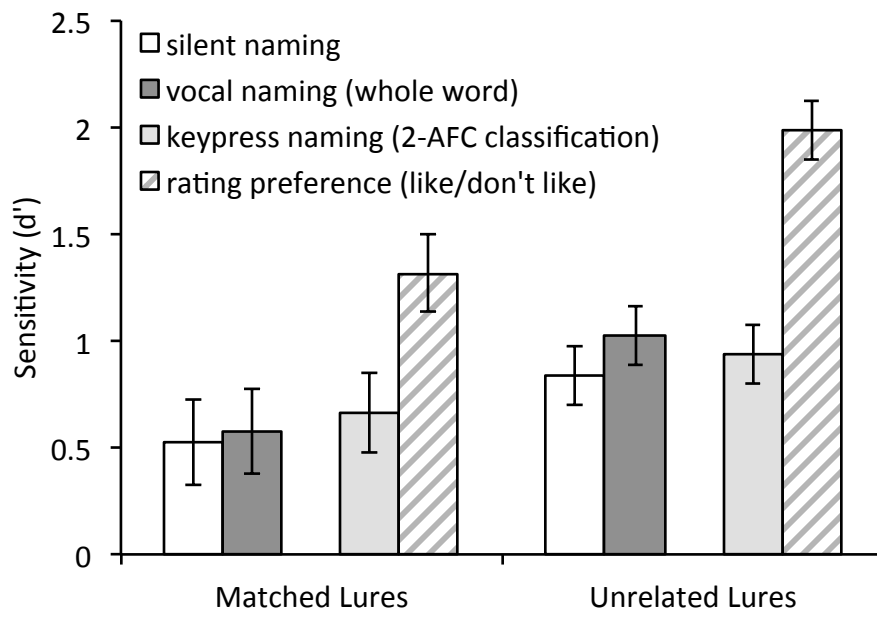


Figure 6