Replication of Experiment 2 by Brady, Konkle, Alvarez, & Oliva
(2013, *Journal of Experimental Psychology: General*)

Karen LaRocque
karenlarocque@stanford.edu

Introduction

Brady and colleagues examined whether features of real-world objects are bound in memory as a single unit or are stored as a collection of independent features. Specifically, they examined whether multiple features of an object are forgotten as a whole (suggesting they are bound as a single unit) or independently. In Experiment 2, the target of this replication, the dependency of memory for exemplar (e.g., which glass) and state (e.g., how full was the glass) of a previously encountered object was assessed. After viewing a series of objects, subjects were asked to select the objects that they had previously viewed from a series of quartets (each containing 2 exemplars x 2 states), and their choices were coded for accuracy of exemplar and accuracy of state. Perfect dependency of feature memory would be demonstrated if participants always either correctly remembered both exemplar and state or remembered neither exemplar or state (after accounting for guessing), while perfect independence of feature memory would be demonstrated if participants jointly remembered both exemplar and state no more frequently than chance. Because the interpretation of object feature dependency during memory retrieval is complicated by attentional fluctuations during memory encoding, the critical measure was not the absolute degree of dependency but whether the degree of dependency decreased with an increasing delay between encoding and retrieval (suggestive of independent forgetting of object features). Using separate short-delay and long-delay groups, Brady and colleagues found that the degree of dependency in memory for object features decreased over a three day delay between encoding and retrieval, providing support for the conclusion that object features are not obligatorily bound together in memory. **This finding of higher rates of feature dependency for the short-delay group relative to the long-delay group is the target finding for the present replication.** Brady and colleagues found that dependency of confidence judgments for the two features was also greater for the short-delay group relative to the long-delay group. However, due to time and budget constraints. this piece of the experiment will be omitted for this replication.

Methods

Power Analysis

In the short-delay group, dependence of memory for object features was significantly greater than 0% (D(state|exemplar) = 46.6%, SEM = 9.7%, Cohen's d = 1.2, D(exemplar|state) = 27.4%, SEM = 4.9%, Cohen's d = 1.45). In the long-delay group, dependence of memory for

object features was not significantly greater than 0% (D(state|exemplar) = 13.4%, SEM = 14.1%, Cohen's d = .27, D(exemplar|state) = 7.6%, SEM = 8.7%, Cohen's d = .24). Most critically, dependence was higher for the short-delay group than the long-delay group (Cohen's d = .75 and .78 for D(exemplar|state) and D(state|exemplar), respectively.

Given an effect size of 1.2 for greater dependence than 0%, a power analysis revealed that a sample size *within each group* of N = 8, 10, and 12 would be required for 80%, 90%, and 95% power in a one-sample t-test, respectively.

Given an effect size of .75 for the difference in dependence across the short-delay and long-delay groups, a power analysis revealed that a sample size *within each group* of N = 29, 39, and 48 would be required for 80%, 90%, and 95% power in a two-sample t-test, respectively.

Planned Sample

We selected our sample size of N = 58 (29 participants in each group) to achieve 80% power in replicating the effect of significantly greater dependence in the short-delay group relative to the long-delay group. This sample size will also allow greater than 95% power to replicate the effect of significantly greater dependence than 0% in the short-delay group. Participants will be recruited on Amazon Mechanical Turk and will be required to be located within the United States, have an 85% work acceptance rate, and be using Chrome as their browser. Brady et al. restricted their sample to ages 18 to 35 but we will be accepting any age group for this replication (and will collect demographic information).

Two participants were excluded from the original report because their answers during the encoding period failed to correlate with the answers of the remaining participants (Tim Brady, personal communication). For the replication we will assess accuracy during the encoding period on a subset of stimuli that were selected to be unambiguous in the size task (20 smaller than a shoebox, 20 larger than shoebox). Participants will not be invited to participate in the retrieval phase if they score below 70% accuracy on either the "smaller-than" subset or the "larger-than" subset during the encoding phase. Participants will also be excluded if they have a score on the memory task that does not significantly differ from chance at p = .05 (two-tailed) as assessed by a binomial test (assuming 120 four-alternative forced-choice trials, less than or equal to 33% accuracy) or fail to return for an experimental session between 60-84 hours (2.5 - 3.5 days) after the first session (regardless of whether they are in the short-delay or long-delay condition). Finally, it is technically possible for participants to read the source code and generate valid codes that they were not given. Any participants who did not complete the encoding task or who altered their codes to complete the retrieval task during a time other than the allotted window will be excluded.

Materials

We used the stimuli described in the original report, presented at 250 x 250 pixels:

"Object images were chosen from previously published sets of stimuli (Brady et al., 2008; Konkle et al., 2010), supplemented with additional images from a commercially available database (Hemera Photo-Objects, Vol. I and II) and Internet searches using Google Image Search. Overall, 120 basic-level categories of object were selected, and for each of these categories we selected two matching state images for each of two category exemplars. This yielded 120 object categories with 4 images each (2 exemplars x 2 states; see Figure 3)."

Procedure

We followed the procedure described in the original report, omitting the confidence judgments and adding the requirement that the short-delay group return after a 3-day delay, as described below.

Original procedure:

"The experiment consisted of a study phase and a test phase. In the study phase, observers were shown 120 objects one at a time for 200 ms each at the center of the display with an 1,800-ms interstimulus interval. During the presentation of the objects, they judged the physical size of the object (whether it was larger or smaller than a particular container they were shown, which was slightly smaller than a shoebox).

Following this task, they were given a surprise long-term memory task, either immediately following the study period (short delay) or after a 3-day delay (long delay). In the long-delay condition, observers were told immediately after the study period they would need to return in 3 days to perform memory tests. We used a surprise memory test and a 3-day delay to ensure that observers' performance was off ceiling at short delay and decreased substantially between the short and long delay, given that previous work has shown observers are quite good at these comparisons even after 5 hours of studying a large number of objects (Brady et al., 2008). To probe which properties of each object were encoded, we presented a four-alternative forced choice test display for each object, consisting of two exemplars (one familiar, one novel), each in two states (one familiar, one novel). Observers used the mouse to click on which of the four images they believed they had seen previously. After choosing an image, they separately reported how confident they felt (high or low) on both the state comparison and the exemplar comparison.* The next trial then began automatically. There was no feedback."

*"... after observers chose their answer, we highlighted two of the objects (the one they chose and the change-of-state object) and they indicated how sure they were that the correct answer was the one they chose and not the other object (low or high confidence); then we did the same for the change-of-exemplar object."

Modifications:

The procedure was modified to accommodate the Amazon Mechanical Turk environment. Participants initially completed only the encoding task. If their performance did not

fall within the exclusion criteria, participants were told that they had qualified to take a memory test for the objects that they had just seen and were provided with a link and a code that would allow them to complete the retrieval test during a given time window (within the next 10 minutes (short-delay group) or 60-84 hours later (long-delay group)). Participants in the long-delay group received an additional email prompt with the link and the code 60 hours after completing the encoding phase. Due to concerns about considerable attrition in the long-delay group in the Amazon Mechanical Turk environment differentially affecting performance in the two groups, we also asked participants in the short-delay group to return after a 3-day delay (defined as 60 - 84 hours after completion of the first session). After completing the retrieval task, participants in the short-delay group were provided with a link and a code that would allow them to complete a final task 60-84 hours. This information was also provided via an additional email prompt with the link and the code 60 hours after completing the retrieval phase. When the short-delay group returned after the 3-day delay they were asked to press a "start" button and then were told that the experiment was over. We excluded any participants in the short-delay group who did not complete this task, thus matching the two groups on propensity to return after a 3-day delay and accompanying covariates.

Participants were paid $.30 for the encoding phase and $.35 for the retrieval phase. Participants in the long-delay group received a $.10 bonus with the email prompt asking them to return to complete the final portion. Participants in the short-delay group received a $.01 bonus with the email prompt asking them to return for the follow-up task and $.10 for completing the final follow-up task after the 3-day delay.

Analysis Plan

We used the exclusion criteria described in "Planned Sample" and followed the analysis plan described in the original report.

We calculated dependence scores for each participant as described in the original report:

"To address our main hypothesis, we examined the level of dependence between observers' reports of the state and exemplar properties. To do so, we calculated how much more likely observers were to get one property correct (e.g., state) if they got the other property correct than if they got it incorrect, taking into account the contributions of random guessing. In order to convert this into a dependence measure (% dependent), we first formalized two models: a fully independent model in which the properties are stored and forgotten independently and a fully bound model in which the properties are always stored and forgotten together. Then, we quantified where our observed data fell in between the predictions of the two models. Finally, for our critical comparison, we examined how this dependence score for the two properties changed between the short and long delays.

In the fully independent model (referred to as $D = 0$ below), there is never any benefit for memory of the state property given that exemplar was remembered, because the two properties are independent. Thus, no

matter what the overall percent correct is, for an independent model of these two properties, the added memory benefit to one of remembering the other is 0:

$$P^{+}_{D=0}\,(state \mid exemplar) = 0$$

In the fully dependent model (referred to as $D = 1$ below), if the exemplar information is remembered, the state information will always be remembered. If all the objects are remembered, the increased memory performance for state information given exemplar information will go from chance (0.5) to remembered (1.0), for a maximal added benefit of 0.50. However, if observers do not remember an object, we assume they guess randomly from among the four items on the test display, and thus this guessing is independent for the two properties. As a consequence, even in the case of a fully bound underlying representation, random guessing for forgotten items will bring the added benefit down from 0.50. To account for this random guessing, we computed the guessing- adjusted fully bound model, based conceptually on that of Gajewski and Brockmole's (2006) model of boundedness in short-term memory, as follows.

First we estimate the percent remembered (R) for each property, based on the overall percent correct:

$$R(pc) = 2pc - 1$$

This formula treats memory as high threshold and takes into account that any overall percent correct (pc) was achieved not only because items were remembered but also because items were sometimes forgotten but guessed correctly (Macmillan & Creelman, 2005). The "adjusted percent remembered" R estimates how often observers truly remember a property, after accounting for fortunate guesses, and is calculated based on overall performance and chance (here, 50% for each property).

For a given percent correct, the expected p^{+} (state | exemplar) according to the bound model can then be calculated: Anytime observers remember the property (R% of the time), they should have complete dependence, p^{+} (state | exemplar) = 0.5, and anytime they forget a property (1 - R% of the time), guessing should cause complete independence, p^{+} (state | exemplar) = 0. Thus, although in theory a fully bound representation would have a p^{+} (state|exemplar) of 0.5, once we take into account guessing, the dependence expected in a fully bound model (referred to as $D = 1$ below) varies as a function of overall percent correct (see Appendix A for derivation and simulation code):

$$p^{+}_{D=1}\,(state \mid exemplar) = R(pc) \,/\, (R(pc) + 1)$$

These expected dependences between properties in a fully independent model and in a fully bound model are plotted in Figure 4 as solid black lines.

Based on these models, for each observer we computed how dependent performance was between the state and exemplar conditions. This number could be a value between 0 (fully independent) and 1 (fully dependent), and it was computed based on the percentage of the way between the independent and bound model predictions the observer's p^{+} ($state$ | $exemplar$) was at the observed percent correct. Because the fully independent model always predicts p^{+} D=0 ($state$ | $exemplar$) = 0, this reduces to simply

$$D = p^{+}(state \mid exemplar) / p^{+}_{D=1}(state \mid exemplar)$$

where D is the dependence score of the observer, p^{+} ($state \mid exemplar$) is how much more likely the observer was to get the state correct if he or she got the exemplar correct, and $p^{+}_{D=1}(state \mid exemplar)$ is the bound model prediction at the observer's percent correct."

As in the original report, we then (a) compared dependence scores within each group (separately for D(state|exemplar) and D(exemplar|state)) to 0% (using two-tailed one-sample t-tests) and (b) compared dependence scores of the short-delay group to the long-delay group (using a two-tailed two-sample t-test).

Differences from Original Study

The original study was conducted in the lab and participants were recruited from the MIT participant pool, whereas the present replication was conducted online and recruited participants from Amazon Mechanical Turk. The use of an online sample may increase variability in attentional fluctuations during encoding. However, this difference is not anticipated to impact the critical qualitative pattern of results (although it may decrease overall accuracy), as the original report emphasizes that the comparing of the short-delay and long-delay groups controls for effects of such attentional fluctuations. The original study collected confidence judgments on every trial reflecting how sure participants were that they had selected the correct exemplar and how sure participants were that they had selected the correct state. The present replication omits these confidence judgments for time and budgetary purposes, with the hope that they do not orient the participants' attention to feature dependency in a way that would alter their forced-choice accuracy dependence scores. Finally, we use an 'opt-in' rather than 'opt-out' procedure (used in the lab) for having participants complete the retrieval test. This may alter the pool of participants who complete the retrieval test relative to the lab study, but we do not expect systematic differences between the short-delay and long-delay groups.

(Post Data Collection) Methods Addendum

Actual Sample

Forty-nine participants completed the encoding phase (mean age of 32.47 (SD = 9.92; range = 19-58; 21 female, 27 male, 1 no gender reported). Of these 49 participants, 43 reached the accuracy criterion and were invited to complete the retrieval task, and 24 of these participants completed the retrieval task (16 short-delay group, 8 long-delay group). Four of the 8 participants in the long-delay group performed at chance (accuracy less than 33%, as specified in exclusion criteria) and thus were excluded (no participants in the short-delay group were excluded for chance performance). Only 2 of the short-delay group participants completed the final check-in, although 3 other short-delay group participants did email the experimenter after three days in regards to the final check-in. Using the criteria specified in our pre-data collection plan, there were 6 participants eligible for analysis (2 short-delay participants, mean

age of 27.5; 4 long-delay participants, mean age of 37.75). Exploratory analyses included the 3 participants in the short-delay group who emailed the experimenter after three days (5 short-delay participants, mean age of 33.0) or all short-delay group participants (16 short-delay participants, mean age of 30.25).

Differences from pre-data collection methods plan

Based on initial piloting, the ISI was changed from 1800 ms to 2000 ms.

Participants were not emailed at exactly 60 hours after they became eligible for the final phase of the experiment, but all participants were emailed with at least 12 hours of time left to complete the final phase of the experiment.

Due to low participation rates, any participants who returned after three days also received an additional $.10 bonus as an extra incentive to return and complete the study.

During analysis t-tests were conducted within a regression framework, and thus the standard error estimates for comparisons of short-delay dependence to chance are slightly different than those in the original study.

Results

Data preparation

Dependency scores for state and exemplar information were computed for each participant following the procedure used by Brady et al., 2013.

Confirmatory analysis

The short-delay and long-delay groups did not differ in dependency scores for state (short-delay M = .53, SE = .43; long-delay M = .37, SE = .30) or exemplar information (short-delay M = .44, SE = .50; long-delay M = .55, SE = .35) , both $t < 1$. Participants in the short-delay condition did not show significantly greater dependency than zero for state or exemplar information, $t(7) = 1.22, .87$, $p = .29, .43$, respectively.

Exploratory analyses

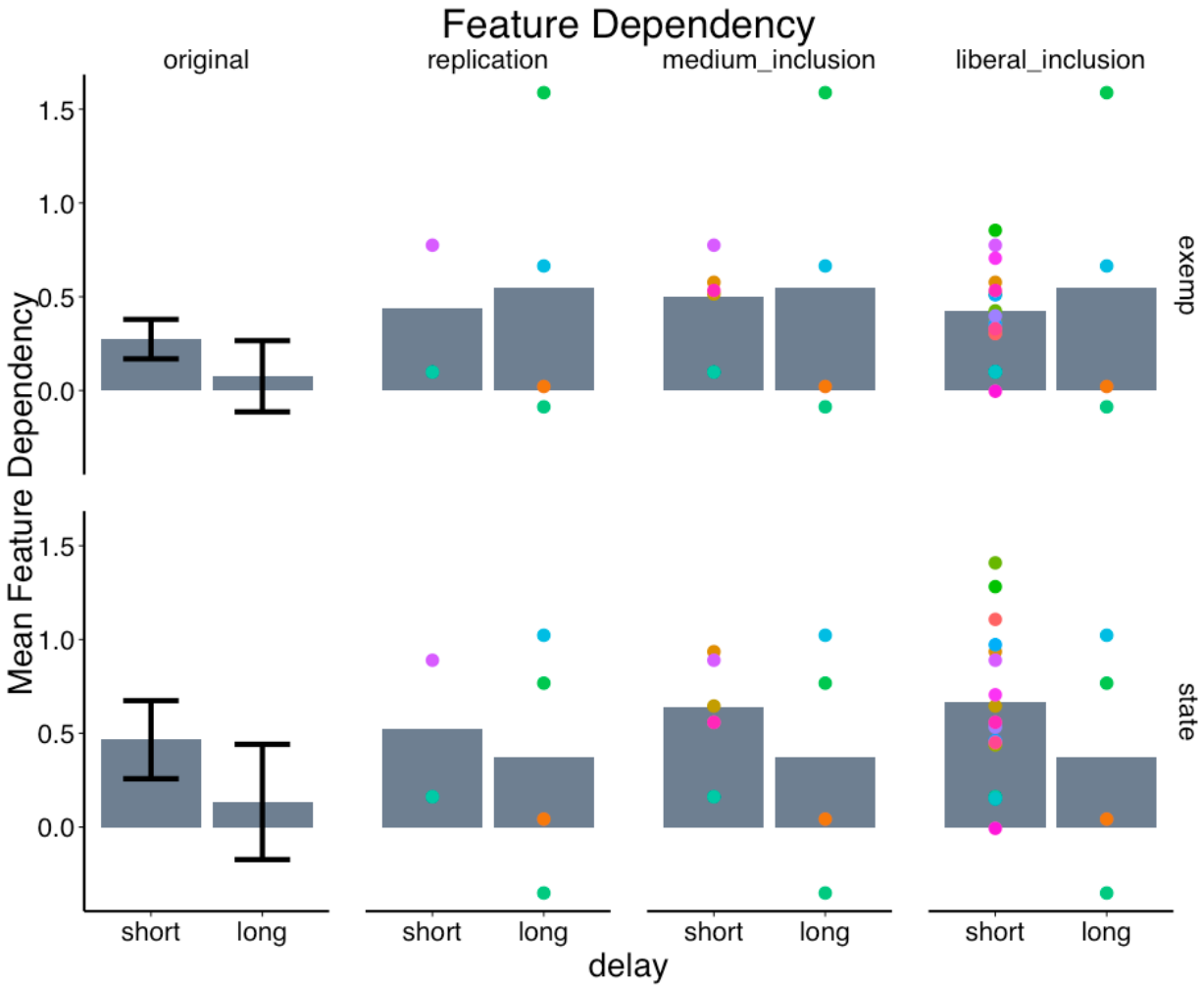
Due to the sample size of n = 2 in the short-delay group, two additional analyses were conducted.

First, the 3 participants who emailed the experimenter regarding the check-in HIT were included in the analysis. Again, the short-delay and long-delay groups did not differ dependency for state (short-delay M = .64, SE = .21; long-delay M = .37, SE = .24) or exemplar information (short-delay M = .50, SE = .24; long-delay M = .55, SE = .27) , both $t < 1$. Participants in the short-delay condition showed significantly greater dependency than zero for state information, $t(7) = 2.99$, $p = .02$ and marginally significantly greater dependency than zero for exemplar information, $t(7) = 2.08$, $p = .08$.

Second, all short-delay participants were included in the analysis. Again, the short-delay and long-delay groups did not differ dependency for state (short-delay M = .67, SE = .11; long-delay M = .37, SE = .23) or exemplar information (short-delay M = .43, SE = .10; long-delay M = .55, SE = .19) , $t = 1.18, .55$, $p = .26, 59$, respectively. Participants in the

short-delay condition showed significantly greater dependency than zero for state information, t(18) = 5.90, p = < .01 and significantly greater dependency than zero for exemplar information, t(18) = 4.47, p = .01.

The mean dependency scores in the original manuscript and in each of the three analysis groups in the present replication are displayed below.



Discussion

Summary of Replication Attempt

Neither the confirmatory analysis nor the exploratory analyses replicated the result of lower feature dependency following a long delay between encoding and retrieval relative to a short delay between encoding and retrieval. The confirmatory analysis did not reveal feature

dependency greater than zero in the short-delay group, but the exploratory analyses with additional participants did replicate this result.

Commentary

Due to a combination of low participation rates in the encoding task, high attrition, and chance memory performance in the long-delay group, this replication attempt was severely underpowered. Post-hoc power analysis revealed power of .10 in the confirmatory analysis and .24 in the exploratory analysis using a liberal inclusion criteria (assuming Cohen's d of .75, from original result). Due to these levels of power, it is not possible to draw any conclusions from the observed null effects and the replication is inconclusive. In addition to the low levels of power due to the small sample size, it should also be noted that (likely due to the small sample size) the two delay groups were not well matched on age.